



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

보건학박사 학위논문

DNA Methylation Changes
as an Exposure Signature
of Cigarette Smoking

흡연 관련 후성유전학 지표 발굴 연구

2019 년 8 월

서울대학교 보건대학원
보건학과 유전체역학 전공
김 은 애

Abstract

Introduction: Mounting evidence suggests that both genetics and environments shape DNA methylation (DNAm) status throughout lifetime. Little is known about reproducible DNAm changes that are specifically induced by environmental exposure. This study aimed to identify exposure-specific DNAm changes, particularly due to smoking. We first investigated smoking-associated DNAm changes in monozygotic (MZ) twins. CpG sites (CpGs) associated with smoking were subsequently examined for possible genetic control by methylation quantitative loci (mQTL). Finally, we evaluated DNAm score using smoking-associated CpGs for prediction of smoking.

Methods: We obtained peripheral blood DNAm data of 385 samples (95 pairs of MZ twins for the discovery set and 195 non-MZ twin first-degree relatives for the validation set) from the Korean Healthy Twin (KHT) using Illumina's HumanMethylation450 array. An additional validation set of 149 samples (61 pairs of MZ twins and their first degree relatives) from the KHT were analyzed using Illumina's Infinium MethylationEPIC BeadChip array. We also obtained peripheral blood DNAm data of 479 individuals (66 pairs of MZ twins for the discovery set and 347 non-MZ twins for the validation set) from the Australian Mammographic Density Twins and Sisters Study (AMDTSS), using Illumina's Infinium HumanMethylation450 array. We tested associations between smoking and DNAm changes across >18,000 CpGs that were previously reported to be associated with smoking. After assessing study-specific smoking-associated CpGs, we meta-

analyzed the two studies. To identify genetic control over DNAm, we performed methylation quantitative loci (mQTLs) analyses using the KHT genotype data of a total of 289 individuals, followed by subsequent examinations of whether those mQTLs are associated with smoking. Finally, we computed weighted DNAm score to assess its performance for prediction of smoking.

Results: In the KHT MZ twins, 8 CpGs were significantly associated with high-dose smoking exposure (≥ 10 pack-years) at the suggestive significance threshold of $p < 5e-5$, including CpGs in *AHRR*, *2q37.1* (in the vicinity of *ALPPL2*), *MYO1G* and *IL*. In the analysis of the AMDTSS MZ twins, 5 CpGs (annotated to *2q37.1*, *VAR5* and *AHRR*) were significantly associated. In the meta-analysis, 14 CpGs in *AHRR*, *2q37.1*, *MYO1G* and *F2RL3* were significantly associated with ≥ 10 pack-years of smoking. In the mQTL analysis, 3,609 (19.7%) of the previously reported $> 18,000$ smoking-related CpGs were significantly associated with at least one proximal SNP (*cis*-mQTL) at Bonferroni-corrected $p < 0.05$. 185 (22.1%) out of the smoking-associated 838 CpGs (meta-analysis of associations $p < 0.05$) were associated with *cis*-mQTLs (Bonferroni-corrected $p < 0.05$). None of the significant mQTLs were associated with smoking. DNAm score based on smoking-associated CpGs ($p < 5e-5$) was computed for prediction of smoking, yielding an AUC of 0.917, 0.895 and 0.84 for the KHT I and II and the AMDTSS validation sets, respectively. With the exclusion of mQTL-associated CpGs (association $p < 0.05$ with smoking), AUC has significantly improved (0.745 to 0.777, 0.7 to 0.734 and 0.61 to 0.646 for the KHT I and II and the AMDTSS validation sets, respectively).

Discussion: We found epigenetic signatures of smoking across multiple loci including *AHRR*, *2q37.1*, *MYO1G* and *F2RL3*. ~20% of the previously reported smoking-associated CpGs were under significant genetic control. DNAm score using the most significant CpGs was informative of predicting high-dose ever-smoking status. CpGs that are independent of effects of mQTLs showed superior performance in predicting smoking. A set of DNAm-associated markers may serve as a stable biomarker of exposure to smoking.

Keywords: Epigenetics, DNA methylation, epigenome-wide association studies, mQTL, Smoking, Biomarker, Monozygotic twin study

Student number: 2015-31282

Contents

Abstract	1
List of Tables	7
List of Figures	9
I. Introduction	11
1. Epigenetics	11
1.1 Overview of Epigenetics	11
1.2 DNA Methylation	12
1.3 Genome-wide DNA Methylation Profiling	13
1.4 Epigenome-wide Association Study (EWAS)	15
2. Epidemiology of Smoking: Health Consequences and Assessment of Exposure to Smoking	17
2.1 Health Consequences of Exposure to Smoking	17
2.2 Assessment of Exposure to Smoking: Biomarkers of Smoking	18
3. Objectives	21
II. Mini-review: Normalization and Cell-type Heterogeneity Deconvolution 23	
1. Normalization	23

1.1	Within-array Normalization.....	24
1.2	Between-array Normalization.....	25
2.	Cell-type Heterogeneity and Deconvolution	26
2.1	Reference-based Cell-type Deconvolution	26
2.2	Reference-free Cell-type Deconvolution	27
2.3	Choice of Cell-type Deconvolution Algorithms	28
III.	Profiling Smoking-Associated DNA Methylation Changes....	30
1.	Material and Methods	30
1.1	Study Design and Population.....	30
1.2	DNA Methylation Data.....	31
1.3	Genotype Data	33
1.4	Genome-wide DNA Methylation Associations with Smoking	34
1.5	mQTL Analysis.....	35
2.	Results.....	38
2.1	Characteristics of the Study Population	38
2.2	Smoking-associated DNA Methylation Changes	39
2.3	Genetic Influences of Smoking-associated DNA Methylation Changes .	43
3.	Discussion	45

IV.	Prediction of Exposure to Smoking Using DNA Methylation	
Score	76	
1.	Material and Methods	76
1.1	DNA Methylation-based Score.....	76
1.2	Assessment of Performance of DNAm-based Score	79
2.	Results.....	81
2.1	DNA Methylation Score by Smoking Status	81
2.2	Prediction of Smoking Exposure using DNA Methylation Score	84
2.3	Improvement of Prediction of Smoking Using DNA Methylation Score by Marker Sets	85
3.	Discussion.....	88
V.	References	113
VI.	Abstract in Korean (국문 초록)	i

List of Tables

Table 1. Studies on assessment of DNAm as a biomarker of smoking	22
Table 2. Methods of normalization and correction for cell-type heterogeneity of DNA methylation data	29
Table 3. Characteristics of the study population of the Korean Healthy Twin (KHT) study.....	51
Table 4. Characteristics of the study population of the Australian Mammographic Density Twins and Sisters (AMDTSS) study.....	52
Table 5. Top smoking-associated CpGs (KHT, $n=190$; pack-year cutoff=10)	53
Table 6. Top smoking-associated CpGs (AMDTSS, $n=132$; pack-year cutoff=10)	54
Table 7. Meta-analyses combining EWAS results of KHTS and AMDTSS	55
Table 8. Differentially methylated regions (DMRs) in relation to smoking.....	56
Table 9. The top traits enriched by differentially methylated sites in relation to smoking (meta-analysis)	57
Table 10. The top ontologies enriched by differentially methylated CpGs in relation to smoking (meta-analysis)	58
Table 11. Top smoking-associated CpGs (meta-analysis) under genetic control (Bonferroni-corrected $p<0.05$).....	59

Figure 11. Associations of mQTLs with smoking exposure	75
Table 12. Performance of DNAm score for prediction of smoking according to marker inclusion thresholds (KHT)	93
Table 13. Improvement of performance of DNAm score as a classifier of smoking according to marker inclusion thresholds (KHT)	94
Table 14. A list of marker sets that yielded the highest AUC values in predicting smoking.....	95

List of Figures

Figure 1. Pipeline for preprocessing DNA methylation data	60
Figure 2. Distributions of all-sample mean DNA methylation values by Infinium probe type (I and II).	61
Figure 3. Distributions of all-sample mean DNA methylation values by processing steps (Raw, normalized and batch-effect removed values)	63
Figure 4. Visualizations of EWAS results of KHT (A. Manhattan plot, B. Q-Q plot, C. Volcano plot)	64
Figure 5. Visualizations of EWAS results of AMDTSS (A. Manhattan plot, B. Q-Q plot, C. Volcano plot)	66
Figure 6. Visualizations of EWAS results of Meta-analysis (A. Manhattan plot, B. Q-Q plot, C. Volcano plot)	68
Figure 7. Comparison of effect sizes and significance levels of EWAS results (KHT vs. AMDTSS).....	70
Figure 8. Dose-response analysis of DNAm levels of cg05575921 and smoking-related dose (KHT)	71
Figure 9. Dose-response analysis of DNAm levels of cg21566642 and pack-years (AMDTSS).....	73

Figure 10. Relationships between significance levels of association between CpG-level DNAm and smoking (x axis) and mQTL significance levels (y axis)	74
Figure 12. Distributions of smoking-associated DNAm score (Pack-year cutoff=10) ($p<5e-5$)	97
Figure 13. Distributions of smoking-associated DNAm score by smoking status (Current/former/never smokers) ($p<5e-5$)	99
Figure 14. Dose-response relationships between dose of smoking exposure and DNAm score	101
Figure 15. ROC curves for DNAm score as a classifier of ever smoking (Pack-year cutoff=10)	108
Figure 16. AUC for prediction of smoking using DNAm score of CpGs selected according to the number of top smoking-associated CpGs and exclusion cutoff of mQTLs (KHT validation set I).....	110

I. Introduction

1. Epigenetics

1.1 Overview of Epigenetics

Epigenetics, literally meaning “on top of” or “in addition to” genetics, is the study on heritable changes to the genome that affect gene expression patterns without changing the DNA sequences. The term “epigenetics” was coined by Conrad H. Waddington to explain possible developmental links between genotypes and phenotypes¹. He proposed a concept “epigenetic landscape”, on which pluripotent cells (likened to a marble on top of the hill) are differentiated through specific pathways (paths) into different outcomes or cell fates (destinations). The cellular decision making process (choices at every endpoint) is governed by genes, while it can be perturbed by the environmental factors.

Recent development of next-generation sequencing and microarray technologies facilitated epigenomic (which refers to the entirety of epigenetic modifications) research. The main focus of epigenomic discipline is the principal mechanisms of epigenetic modifications including DNA methylation (DNAm), histone modifications and regulation by non-coding RNAs (ncRNAs). DNAm is a process in which a methyl (CH₃-) group is covalently added to the DNA molecule (details of DNAm can be found in the following Chapter I.1.2). Histone modification is post-translational modification of histone proteins in the forms of methylation,

phosphorylation, acetylation, ubiquitylation, and sumoylation. ncRNAs, such as miRNA, siRNA, piRNA and lncRNA, are involved in regulating gene expression.

Epigenome is largely characterized by two opposing features, plasticity and stability. Plastic nature allows for changes during development and in response to environmental stimuli, while such epigenetic changes are potentially reversible. Epigenetic marks, once modified, can be stably maintained throughout the somatic cell divisions. Epigenetics has been thus highlighted for its possible applications as biomarkers for exposure to several environmental factors and health status or diseases^{2,3}.

1.2 DNA Methylation

DNA methylation (DNAm) is among the best studied and most understood epigenetic mechanisms. Most of DNAm in the human genome (70-80%) occurs in the form of covalent addition of methyl (CH₃-) group to C-5 position of cytosine residues, producing 5-methylcytosine (5mC)⁴, which is also informally referred to as the fifth base⁵. It typically occurs in cytosines of the CpG (cytosine-phosphate-guanine) context. CpG islands, mostly residing within promoter regions, are hypomethylated, which is strongly associated with activating genes. Meanwhile, CpGs located within intergenic or repetitive regions are mostly hypermethylated, most of which activity is associated with repressing gene expression to maintain genomic stability.

DNAm is established and maintained by DNA methyltransferases (DNMTs) with the help of methyl-CpG binding proteins (MBDs). DNMT3a/b are responsible for *de novo* DNA methylation. Once DNAm marks are established, DNMT1 is involved in maintenance methylation activity, resulting in mitotically heritable changes in DNAm.

A growing body of epigenome-wide association studies (EWAS) have indicated that DNAm can be used as biomarkers for exposure and risk prediction, early detection and prognosis of diseases⁶. Notably, DNAm changes are induced by age and a range of environmental factors throughout lifetime, including intrauterine/early-life environmental factors⁷⁻¹⁰, exposure to chemicals¹¹⁻¹³ and lifestyle-related factors¹⁴⁻²². Some of these exposure-induced DNAm changes consequently result in various health effects^{2, 3, 6, 7, 23, 24}, by altering gene expression levels. DNAm is also influenced by intrinsic factors, that is, genetic sequence variations known as methylation quantitative loci (mQTLs)²⁵⁻²⁷.

1.3 Genome-wide DNA Methylation Profiling

The advent of Illumina's microarray-based DNAm assays has accelerated human DNAm studies. There are several platforms that measure DNAm of targeted regions such as bisulfite pyrosequencing²⁸, MethyLight²⁹ and EpiType³⁰. Meanwhile, Illumina has sequentially released three different assay platforms for comprehensive scanning of DNAm levels across the genome, including HumanMethylation27

(27K), HumanMethylation450 (450K) and MethylationEPIC (EPIC) BeadChips, each measuring ~27K, ~450K and ~850K DNAm sites, respectively, across the human genome. Such genome-wide DNAm profiling technologies facilitated accumulation of EWAS studies.

As an extension of the 27K chip based on one array that employed the Infinium I probe chemistry only, a set of Infinium II probes were added on the 450K, comprising two different probes (Infinium I and II probes). It was made necessary to apply normalization methods to reduce technical biases due to heterogeneity between the two array types. Between-array normalization methods are elaborated in detail in the following **Chapter II**.

The Infinium I assay has two-color beads per probe, one for the methylated allele in the red channel and one for the unmethylated allele in the green channel. The Infinium II assay uses a single bead per probe in the red or green channel to measure methylated and unmethylated signals, respectively. The intensities of the methylated (M) and unmethylated (U) probes are measured at a single CpG position. DNA methylation levels (β) are computed as

$$\beta = \frac{M}{M+U+\alpha} \quad (\text{Equation 1})$$

which is the ratio of the methylated intensities (M) to the overall intensities (M+U). α , usually set to 100, is a constant for regularizing β when both M and U are small. The raw .idat files that contain the per-probe intensity information for each

sample are loaded and converted into a matrix of β values used for the downstream analyses using software tools designed for DNAm data processing.

1.4 Epigenome-wide Association Study (EWAS)

In parallel to genome-wide association studies (GWAS) scanning associations between single nucleotide polymorphisms (SNPs) and complex traits, EWAS is another genetic epidemiology branch to identify genome-wide epigenetic changes in association with a phenotype of interest. The primary scope of EWAS is at the level of DNAm, given other epigenetic fields such as chromatin and RNA are complex and experimental methods are not well standardized. The Illumina microarray-based DNAm profiling is the most widely adopted experiments for EWAS.

Study design and samples are determined by the purpose/hypothesis of the study. One of the most widely employed study designs is case-control studies using samples nested within the existing cohorts that have been biobanked and deeply phenotyped. Like other observational studies, EWAS can suffer from many forms of biases including publication, ascertainment and selection biases. Several strategies have been attempted to combat such biases and thus make valid causal inference. Compared to cross-sectional studies, longitudinal analyses using the prospective cohort provide robust evidence, particularly for the development or progression of diseases. Monozygotic (MZ) twin studies may also provide compelling evidence of

DNAm changes that are specifically induced by exposure of interest, as they have reduced effects of DNA sequence variation on DNAm^{31, 32}.

Another consideration in the study design is determining target tissues, which can differ by the purpose of studies or the accessibility of tissues if they can serve as a surrogate tissue for the tissue of interest. Blood tissues, commonly deposited in biobanks for its wide utility as a biomarker, can be a surrogate for other tissues such as adipose³³ or brain tissues³⁴, though it requires special attention when analyzing and interpreting results. As some of the DNAm patterns are distinctive of specific tissue or cell types, such heterogeneity can cause possible chances of biases in interpreting EWAS results. Experimental techniques of cell sorting methods such as purified by fluorescence-activated cell sorting (FACS) technology can be used to quantify proportions of blood cell types. Alternative cost-efficient computational methods³⁵⁻⁴⁰ have been developed to address confounding effects due to cellular heterogeneity as reviewed in the following **Chapter II**.

2. Epidemiology of Smoking: Health Consequences and Assessment of Exposure to Smoking

2.1 Health Consequences of Exposure to Smoking

Cigarette smoke, a mixture of thousands of chemicals containing carcinogens such as polycyclic aromatic hydrocarbons (PAH) and N-nitrosamines, causes adverse health effects. Despite being an established modifiable risk factor, smoking affects a large number of preventable deaths in the world, posing great public health burden worldwide. In 2015, 11.5% of global deaths were attributable to smoking and smoking was one of the most important risk factors based on disability-adjusted life-years (DALYs), according to the large-scale systematic review by the Global Burden of Diseases, Injuries, and Risk Factors Study (GBD)⁴¹. The major smoking-induced diseases include coronary heart disease, chronic obstructive pulmonary disease, and cancer including lung cancer and upper aerodigestive tract cancer.

In South Korea, smoking rate was 22.3% (males: 38.1%, females: 6.0%), according to the Korean National Health and Nutritional Examination Survey (KNHANES) 2017, with the male smoking rate ranked among the highest of the OECD countries. The study on population attributable fractions (PAF) of smoking on cancer in South Korea reported that smoking contributed to 20.9% of cancer incidence and 32.9% of deaths due to cancer for men and 2.1% of cancer incident cases and 5.2% of cancer deaths for women in 2009⁴². Of the lung cancer deaths in South Korea, 71% were attributable to smoking⁴². Lung cancer was the 6th most

common type of cancer in South Korea with the prevalence of 4.4% (Men: 6.2%, women: 3.0%), according to National Cancer Statistics Korea 2016. Despite the relatively moderate prevalence among all cancer types, the 5-year relative survival of lung cancer was 28.2% in 2012-2016 (Men: 23.7%, women: 38.6%), which was the second lowest following pancreatic cancer.

2.2 Assessment of Exposure to Smoking: Biomarkers of Smoking

Objective assessment of exposure to cigarette smoke by active smoking is critical in studying smoking-associated health effects. Self-report based assessment such as investigating smoking status (former/never/current smokers) or pack-years ((packs smoked per day) \times (years smoked)) is commonly used, due to its convenience and cost-effectiveness. Assessment of smoking based on self-report shows high accuracy in many epidemiological studies⁴³⁻⁴⁵. However, there are several populations with high disagreement rates between self-report and objective measures of smoking, including females in some East Asian and Islamic countries^{46, 47}, pregnant women⁴⁸ and adolescents^{49, 50}. Of note, self-report assessment of smoking in South Korea, where smoking is perceived negatively, showed high disagreement with urinary cotinine levels in females, which may have contributed to the underestimation of female smoking rates⁴⁶.

There are several alternative biomarkers to quantify exposure to smoking. Exhaled carbon monoxide is used, but detectable for 3-4 hours after last use of

cigarette smoke. Cotinine, a predominant metabolite of nicotine that can be measured in blood, urine, and saliva, is the most widely used biomarker for smoking. Though it shows great usefulness in verifying current smokers, the use is limited to assessing short-term smoke exposure due to its short half-life (16-19 hours), posing its limitations as a biomarker for long-term past exposure.

Mounting evidence of EWAS indicates that DNAm changes can occur in response to exposure to cigarette smoke across multiple genomic regions. Thousands of chemicals contained in cigarette smoke are inhaled into body, affecting multiple tissues including those of lungs and cardiovascular systems. Such damages alter epigenetics of those tissues, which induce inflammation, immune response and impaired vascular functions. DNAm changes in response to cigarette smoking were observed at >18,000 CpGs including *AHRR*, *F2RL3*, *MYO1G* and other genomic loci according to the large-scale meta-analysis¹⁸. Such genes could mediate detoxification of chemicals involved in the smoking-associated metabolisms⁵¹. One of the best-replicated associations is hypomethylation (*i.e.*, decreased DNAm) in the *AHRR* (Aryl-Hydrocarbon Receptor Repressor) region. PAHs induced by smoking activate the arly hydrocarbon receptor (AHR)⁵², in which process hypomethylation of CpGs in the *AHRR* region is followed by increased expression of *AHRR*. Of particular note, smoking-induced DNAm signatures can reflect past as well as current exposure and exposure dosage of cigarette smoking. Previous studies have attempted to evaluate epigenetic assessment of smoking exposure. Several studies reported that a single CpG within the *AHRR* loci,

cg05575921, provided excellent performance in predicting smoking status with an area under the curve (AUC) of over 0.9⁵³⁻⁵⁵. cg05575921 showed comparable performance with serum cotinine levels in discriminating current vs. never smokers, while the DNAm site was more informative in predicting former vs. never smokers compared to cotinine levels⁵⁵. Moreover, combination of multiple smoking-associated CpGs was useful in predicting smoking. Studies on assessment of DNAm as a biomarker of smoking are summarized in **Table 1**⁵³⁻⁵⁶.

3. Objectives

In this study, we aimed to investigate DNAm markers that can reflect changes specifically induced by smoking. We explored genomic and epigenetic landscape of smoking behavior, dissecting effects of genetic polymorphisms and DNAm on smoking using MZ twins, with the following approach. In the first part of the study, we replicated a list of smoking-related DNAm sites using MZ twins, which may reduce possible effects due to population stratification and genetic polymorphisms. Second, we performed mQTLs analyses, where DNAm levels across >18,000 sites were tested for association with proximal single nucleotide polymorphisms (SNP) out of a total of >4 million markers, to examine possible genetic control. Finally, we evaluated smoking-associated DNAm markers as a quantitative method to predict smoking exposure.

Table 1. Studies on assessment of DNAm as a biomarker of smoking

Studies	Study population	DNAm quantification methods	Main results (Prediction ability in AUC)
Shenker, Natalie S., <i>et al.</i> (2013)	Turin component of the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) cohort - Test set: 81 healthy individuals (33 never, 30 former & 18 current smokers) - Validation set: 180 healthy women (102 never, 45 former & 33 current smokers)	Pyrosequencing - cg05575921, cg06644428, cg21566642, cg06126421 - DNAm score computed using $\beta_1M_1 \times \beta_2M_2 \times \beta_3M_3 \times \beta_4M_4$	Never vs. Former smokers - DNAm score: 0.83 (95% CI: 0.70-0.96)
Philibert, Robert, <i>et al.</i> (2015)	National Institutes of Health study that examined the effects of alcohol on DNA methylation - 61 individuals (35 non- & 26 smokers; predominantly white)	Illumina's HumanMethylation450 BeadChip (HM450K) - cg05575921, cg01940273, cg21566642, cg05951221, cg23576855	Recent users of cigarettes vs. non-smokers - cg05575921: 0.99 - cg01940273: 0.94 (+cg21566642: 0.939) - cg21566642: 0.905 (+cg05951221: 0.918) - cg05951221: 0.903 (+cg23576855: 0.919) - cg23576855: 0.860
Zhang, Yan, <i>et al.</i> (2016)	A subset of 1000 participants from ESTHER study (a population-based cohort study conducted in Saarland, Germany) - Discovery set (500 individuals) - Validation set (500 individuals)	HM450K - cg05575921, cg05951221, cg02451831, cg06126421 - DNAm score computed using $\beta_1M_1 + \beta_2M_2 + \beta_3M_3 + \beta_4M_4$	Current (Former) vs. never smokers - cotinine: 0.96 (0.54) - cg05575921: 0.96 (0.78) - cg05575921 and cotinine: 0.98 (-) - DNAm score: 0.97 (0.83) - DNAm score and cotinine: 0.98 (-)
Andersen, Allan M., <i>et al.</i> (2017)	The Family and Community Health Study (FACHS) & The Iowa Adoption Studies (IAS) - FACHS: 592 individuals (346 never, 101 former & 145 current smokers; predominantly European-American) - IAS: 209 individuals (137 never, 16 former & 56 current smokers; predominantly African-American)	ddPCR implementation (quantitative PCR approach) - cg05575921	Smokers vs. non-smokers - cg05575921: 0.897 (95% CI: 0.848-0.947) - Self-report: 0.889 (95% CI: 0.840-0.938) - Self-report and cg05575921: 0.929 (95% CI: 0.884-0.973)

II. Mini-review: Normalization and Cell-type Heterogeneity Deconvolution

This mini-review is dedicated to two major analytical challenges to remove confounding biases in DNA methylation analyses: (1) data normalization and (2) cell-type heterogeneity deconvolution. Data normalization process reduces effects of technical artifacts caused by experimental designs of DNAm assay chip technology. Cell-type heterogeneity deconvolution addresses effects of cellular heterogeneity of DNAm data. The methods developed to address the challenges are summarized in **Table 2**.

1. Normalization

Normalization is a key preprocessing step for analyses of microarray-based chips to remove technical biases caused by microarray technology. Normalization methods for DNA methylation data are largely focused on the Illumina's Infinium platforms, the most widely used technologies to profile genome-wide DNA methylation. The Infinium arrays utilize two different probes (Infinium I and Infinium II probes), which makes it necessary to develop normalization methods that can be specifically used for microarray-based DNA methylation data (The details of the array design are elaborated in **Chapter I**). Scores of normalization methods are currently available; choice of proper normalization methods is required to minimize technical

biases and thus improve data quality, which can influence the results of downstream statistical analyses.

1.1 Within-array Normalization

For within-array normalization, background needs to be corrected to estimate the true signal from the total observed fluorescence signal ('background correction'). The total observed intensity (X_t) is defined as the sum of the true signal (X_s) and the background signal (X_b) ($X_t = X_s + X_b$). The use of two-color beads of the Infinium II allows more loci to be tested on the array simultaneously; however, the platform design leads to a higher background in Infinium II, because the measurement of one of the two beads is affected by the residual emission of the other. The background of Infinium II results in a systematically different distribution of β values between Infinium I and Infinium II, with a less dynamic range of β values than the Infinium I. The methods of background corrections include (1) subtraction of background estimates from negative control probes, as implemented in the '*lumi*' Bioconductor package⁵⁷, and (2) the use of out-of-band (OOB) probes for estimating background using convolution methods (normal-exponential using out-of-band probes, '*noob*')⁵⁸, which is available in the '*minfi*' Bioconductor package⁵⁹.

1.2 Between-array Normalization

Between-array normalization of Infinium DNA methylation data is to remove array-to-array technical biases. The data from Type I and Type II arrays need to be incorporated for the downstream analyses, after normalizing the data sets between arrays. As the Infinium I data are more stable and reproducible compared to the Infinium II data, most normalization methods were developed to reduce biases of Infinium II probes. Most widely used normalization methods include (1) peak-based correction (PBC)⁶⁰ (2) subset-quantile within array normalization (SWAN⁶¹) and (3) beta mixture quantile (BMIQ) normalization⁶². PBC rescales the modes of β values of Type II probes to those of Type I probes. SWAN uses a subset of probes for each probe type that are biologically similar on the basis of CpG content. BMIQ normalization adjusts the distribution of β values of Type II probes to that of Type I probes, using a three-state beta-mixtures model to assign probes according to methylation states (unmethylated/partially methylated/fully methylated).

2. Cell-type Heterogeneity and Deconvolution

DNA samples can be extracted from many different tissues to measure DNAm, depending on the purpose of studies, the type of affected tissues or tissue availability. Given that DNAm is highly cell-type specific^{63, 64}, it is critical to infer cell type heterogeneity and correct for such heterogeneity to address potential confounding factors⁶⁴. DNAm in peripheral blood leukocytes is most widely studied, as it is easily available while representing DNAm of metabolically relevant tissues. Blood tissues are highly heterogeneous, as different blood cell subtypes have differential DNAm profiles and thus their heterogeneity needs to be accounted for. Cell-type deconvolution algorithms are largely categorized into two main paradigms, according to whether DNAm reference profiles are used or not: (1) reference-based and (2) reference-free methods. The choice of appropriate cell-type deconvolution methods needs careful attention.

2.1 Reference-based Cell-type Deconvolution

For reference-based cell-type deconvolution, reference profiles of DNAm of the studied tissues should be constructed in advance. Reference DNAm profiles are constructed over a set of CpGs, which are differentially methylated and thus are representative of each cell type. Reference-based methods are supervised methods that estimate cell type fraction by using (1) constrained projection (CP)³⁶ or (2) non-constrained approach³⁸. The CP approach uses the estimated regression coefficients

on which normalization constraints are imposed. The regression coefficients correspond to the estimated cell-type composition for each sample. This method is widely known as the Housman's method³⁶, implemented in the R package, '*minfi*'. Meanwhile, the non-constrained approach is to first estimate regression coefficients and then to impose normalization constraints on the coefficients. This method is available in the CIBERSORT software³⁸.

2.2 Reference-free Cell-type Deconvolution

Reference-free methods correct for cell-type heterogeneity without prior knowledge of reference profiles of cell types. These methods infer cellular heterogeneity by using (1) surrogate variable analysis (SVA)³⁷ and variants of SVA, such as RefFreeEWAS⁶⁵, independent surrogate analysis (ISVA)³⁹ or removing unwanted variation (RUV)³⁵, or (2) methods adapted within the EWAS framework, which perform EWAS after removing cellular heterogeneity, such as EWASher⁴⁰ or ReFACTor⁶. SVA methods were initially developed to address confounding variations in a more general context, not limited to handle cellular heterogeneity in DNAm data. SVA identifies surrogate variables (SVs) associated with confounding variations that are unrelated to the phenotype of interest.

The key assumptions of algorithms differ by relative data variation. SVA/ISVA algorithms assume that the main variation of the data is associated with the phenotype of interest, not cell-type heterogeneity. Meanwhile, their counterparts,

EWASher and ReFACTor, assume that the data variation is mainly driven by cell-type composition. Therefore, it requires careful attention to choose appropriate methods.

2.3 Choice of Cell-type Deconvolution Algorithms

The selection of cell-type deconvolution algorithms depends on the following⁶⁶.

First, the choice of algorithms can be determined by the purpose of studies. Of note, when researchers need absolute or relative estimates of cell type fractions by cell types within each sample, reference-based approach may be most appropriate, if the reference DB exists. It applies, for example, when identifying cell types harboring differentially methylated sites or characterizing DNAm patterns of specific cell types. Researchers should also examine whether the sound reference database of the underlying tissues is available. Lastly, the proportion of data variance can be critical for selecting algorithms. Selection of algorithms can differ by whether the variance of cellular heterogeneity is larger or smaller than that of the phenotype of interest or that of potential confounding factors.

Table 2. Methods of normalization and correction for cell-type heterogeneity of DNA methylation data

Major analytical challenges for DNA methylation analysis	Methods	Implementations
Normalization		
<i>Within-array normalization (Background correction)</i>	Background subtraction using negative control probes (lumi)	R Bioconductor <i>lumi</i>
	Convolution using out-of-band control probes (noob)	R Bioconductor <i>minfi</i>
<i>Between-array normalization</i>	Peak-based correction (PBC)	R Bioconductor <i>ChAMP</i>
	Subset-quantile within array normalization (SWAN)	R Bioconductor <i>minfi</i>
	Beta mixture quantile normalization (BMIQ)	R Bioconductor <i>wateRmelon</i>
Correction for cell-type heterogeneity		
<i>Reference-based cell-type deconvolution</i>	Constrained projection (CP) approach (Houseman's method)	R Bioconductor <i>minfi</i>
	Non-constrained approach	CIBERSOFT
<i>Reference-free cell-type deconvolution</i>	Surrogate variable analysis (SVA) and variants of SVA (ISVA/RUV)	R Bioconductor <i>sva</i> , <i>isva</i> , <i>ruv</i>
	Implementations within EWAS frameworks	EWASher, ReFACTor

III. Profiling Smoking-Associated DNA Methylation Changes

1. Material and Methods

1.1 Study Design and Population

1.1.1 The Korean Healthy Twin (KHT) Study

A subset of 534 individuals from the Korean Healthy Twin Study (KHT) were included for DNAm measurement experiments. The Korean Healthy Twin Study is a nation-wide twin and family registry that recruited >3,000 individuals and details of the study protocols are provided in Sung J et al. (2006)⁶⁷ and Gombojav B et al. (2013)⁶⁸. The DNAm study set consists of 160 families with MZ twins and their first-degree relatives. Of the 160 families included for analysis, 86 families consisted of MZ twins only, 71 families consisted of both MZ twins and their first-degree relatives and 3 families consisted of non-MZ-twin first-degree relatives. For assessing smoking-associated DNA methylation changes, MZ twins (95 pairs; 190 individuals) were included (Dataset 1: Discovery set). Dataset 2 and 3 were used as validation sets to assess DNAm score computed using smoking-associated CpGs as a predictor of smoking in **Chapter IV**.

1.1.2 The Australian Mammographic Density Twins and Sisters Study (AMDTSS)

The AMDTSS is an Australia-based twin and family study which was initially designed to study mammographic density as a risk factor for breast cancer⁶⁹. The AMDTSS set ($n=479$) consists of 132 female MZ twins and their 215 sisters from 130 families. 66 MZ twins (132 individuals) were used for epigenome-wide association analysis (Dataset 1: Discovery set). 347 non-MZ twins were used as a validation set in **Chapter IV**.

1.2 DNA Methylation Data

1.2.1 KHT

We performed a series of three separate sets of experiments measuring 192 (Dataset 1), 200 (Dataset 2) and 150 samples (Dataset 3), respectively. For DNAm measurements, Illumina's HumanMethylation450 BeadChip (HM450K) arrays were used for the first two experiment sets (Dataset 1 and 2) and Illumina's Infinium MethylationEPIC BeadChip (EPIC) arrays for the Dataset 3.

For three datasets, genomic DNA was extracted from the peripheral blood leukocytes of the study samples. DNAm data preprocessing pipeline is presented in **Figure 1**. CpGs with a detection p -value >0.01 , outside of CpG context or located at sex chromosome were excluded from analysis. For the first two datasets, 466,687

and 464,637 CpGs were retained from the original probe set, respectively, and a consensus set of 462,980 CpGs were included for the further analyses. Samples with detection p -values >0.01 were excluded ($n=2$). Three additional samples were excluded based on predicted gender probability (mismatch between predicted and actual genders). To verify the sample exclusion, we compared their genome-wide DNAm levels with those of other family members, resulting in showing extreme similarities or dissimilarities given the kinship. Beta-mixture quantile (BMIQ) normalization⁶² was performed for each of the data sets separately. BMIQ-normalized beta values of both data sets were then corrected for the known batch effects ('Sentrix ID' that contains positional information in the batch) using the ComBat function⁷⁰. This methodology uses parametric empirical Bayes frameworks for correcting for batch effects, implemented in the 'sva' R package⁷¹. We estimated cell type compositions using the Houseman's reference-based approach³⁶.

For samples of the Dataset 3, DNAm levels at 866,895 CpGs were initially measured using Illumina's EPIC arrays. CpGs with a detection p -value >0.01 , outside of CpG context or located at sex chromosome were excluded from analysis, leaving 821,470 CpGs for the final data set. One sample with a detection p -value >0.01 was removed, leaving 149 samples. After the probe/sample-level QC, BMIQ normalization was followed. This data set was used as a validation set for validating DNAm score as a predictor of smoking in this study (**Chapter IV**).

1.2.2 AMDTSS

Genomic DNA was extracted from dried blood spots. DNAm was measured using the HM450K chips. Raw intensity data were processed using Bioconductor *minfi* package⁵⁹, with which Illumina's reference factor-based normalization methods and subset-quantile within array (SWAN) normalization methods⁶¹ were applied. CpGs with detection p -values > 0.01 or at sex chromosomes were excluded, leaving 468,406 CpGs for analyses. A consensus set of 459,705 CpGs that overlap with the preprocessed KHT data sets were included for the downstream analyses.

1.3 Genotype Data

To examine possible effects of genetic polymorphisms on DNAm, we performed mQTL analysis. Genotype data of the KHT were available for all of the $> 3,000$ study participants (for MZ twins, one genotyped per pair) and we used the subset of genotype data for the individuals for whom DNAm analyses were conducted. We included from the KHT one per each of the 95 MZ twin pairs ($n=95$) and their first-degree relatives ($n=194$) for whose DNAm is measured. Genomic DNA was extracted from whole blood samples of the study population ($n=3,474$). Genotyping was performed using two platforms, Affymetrix Genome-Wide Human SNP array 6.0 (Affymetrix, Inc. Santa Clara, CA, USA) ($n=2,260$) and Illumina Infinium HumanCore-24 BeadChip (Illumina, San Diego, CA, USA) ($n=1,194$).

Markers were filtered out using the PLINK software⁷² based on the following criteria: minor allele frequency (MAF)<0.01, duplicated SNPs, deviation from Hardy-Weinberg Equilibrium (HWE) (p-value<1e-6) or genotype call rates<0.9. A total of 7,422,259 imputed markers that meet information (INFO) score>0.6 in either of the data sets were included. For the 385 individuals included for the DNAm analysis, markers with MAF>0.05, genotype call rates>0.9 and HWE p-value>1e-5 were included for the further analyses ($n=4,098,302$). The KHT genotype data were used for mQTL and genome-wide association tests of smoking.

1.4 Genome-wide DNA Methylation Associations with Smoking

We analyzed 18,496 CpGs for KHT and 18,438 CpGs for AMDTSS that overlap with a list of 18,760 CpGs previously reported to be associated with smoking in the large-scale meta-analysis by Joehanes R *et al.* (2016)¹⁸. DNAm levels of MZ twins were compared within each pair after adjusting for age, sex, BMI and cell type compositions using the empirical Bayes method in the limma framework⁷³.

In addition to identifying CpG site-level associations, differentially methylated region (DMR) analysis was performed using the DMRcate method⁷⁴ as implemented in the ‘*DMRcate*’ R package. DMRcate computes a kernel estimate against a null comparison to identify DMRs. A consensus set of 18,439 markers of the two studies, KHT and AMDTSS, were included for meta-analyses. A random-

effects model was fitted to address heterogeneity between the two studies using the restricted maximum likelihood estimation^{75, 76} as implemented in the ‘*metafor*’ R package⁷⁷.

For those CpGs differentially methylated in relation to smoking exposure, dose-response analysis was performed. Restricted cubic spline regression⁷⁸ was used to identify dose-response relationships between DNAm levels and dose-related variables of smoking exposure (pack-years, smoking intensity which is defined cigarettes consumed per day and smoking duration (years)). The analysis was performed using the R packages ‘*rms*’⁷⁹ and ‘*Hmisc*’⁸⁰.

Finally, enrichment analysis was performed using the EWAS knowledge base, EWAS Atlas⁸¹, which integrates ~330K high-quality EWAS associations from 1830 cohorts and 390 ontology entities. Using the list of CpGs provided, significantly enriched traits and ontology terms are obtained. The co-occurrence probability between input CpGs and trait-associated CpGs was calculated using the weighted Fisher’s exact test. The weight of each CpG is defined as the number of studies that reported the corresponding CpG-trait association. Ontology entities are represented in Experimental Factor Ontology (EFO)⁸².

1.5 mQTL Analysis

For mQTL analysis, we fitted a linear mixed polygenic model for each of all SNP-CpG pairs. We estimated polygenic effects implemented the ‘*GenABEL*’⁸³ R

package for each of the >4M SNPs on DNAm levels across >18,000 CpGs after accounting for age, sex, BMI, smoking status and cell type compositions as fixed effects and their familial relatedness as a random effect using a pedigree-based kinship matrix.

After assessing polygenic effects, we used different multiple-testing correction thresholds for markers tested within each of the different window sizes (50kb, 100kb, 500kb, 1Mb and genome-wide), in order to identify *cis*- or *trans*-regulatory effects. Bonferroni-corrected *p*-value thresholds were used for multiple-testing correction. For *cis*-mQTLs, we used uniform expected thresholds for a given window size that correspond to $0.05/(\text{total number of genomic markers after pruning} \times 2 \times \text{window size}) \times \text{genome size}$. The total number of genomic markers after pruning and genomic size correspond to 313,430 and 3,234,830,000, respectively. Linkage disequilibrium(LD)-based SNP pruning was conducted to accounting for LD structures of SNPs, which produced a subset of 313,430 independent markers out of 4,098,302. The thresholds for markers tested within the window sizes of 10kb, 50kb, 100kb, 500kb, 1Mb and whole genome corresponded to 2.58e-2, 5.16e-3, 2.58e-3, 5.16e-4, 2.58e-4 and 5e-8, respectively.

We further assessed whether mQTLs are associated with smoking exposure. A linear-mixed statistical model (LMM) was constructed to test genome-wide associations between smoking and each of the single nucleotide polymorphisms (SNPs) using GenABEL⁸³. For LMM modelling, adjustment was made for age and

sex as fixed effects and their familial relatedness as random effects using a pedigree-based kinship matrix provided.

2. Results

2.1 Characteristics of the Study Population

The characteristics of the study population are presented in **Table 3**. The mean age of the study population was lower in the KHT (Dataset 1: 48.2 years, Dataset 2: 53.8 years and Dataset 3: 39.5 years), compared to that of the AMDTSS (Dataset 1: 55.6 years and Dataset 2: 56.7 years). The sex distribution of the KHT was mostly balanced between males and females across all of the subsets (50.5, 49.2 and 55% for the Dataset 1, 2 and 3, respectively), while the AMDTSS consists of female participants only. The average body mass index (BMI) was lower in the KHT population (24.3, 24.4 and 23kg/m²), compared to that of the AMDTSS population (26.5 and 27 kg/m²).

The KHT Dataset 1 ($n=190$) consisted of 75 (58 current and 17 former smokers) ever-smokers. Of 95 MZ twin pairs, 31 pairs were discordant for smoking status categorized into current/former/never smoker group. 20 MZ twin pairs were discordant for ≥ 10 pack-year ever-smoking status. For the KHT Dataset 2 ($n=195$), 76 (41 current and 35 former smokers) individuals were ever-smokers. Of 76 ever-smokers, 57 individuals consumed 10 pack-years for their lifetime. For the KHT Dataset 3 ($n=149$), 64 (37 current and 21 smokers; current or former smoking status missing for 6 individuals) individuals were ever-smokers. Of 61 MZ twin pairs, 13 pairs were discordant for smoking status. For the AMDTSS Dataset 1, 49 (12

current and 37 former smokers) individuals were ever smokers. Of 66 MZ twin pairs, 23 pairs were discordant for smoking status.

2.2 Smoking-associated DNA Methylation Changes

For DNAm analyses, the KHT dataset 1 and 2 needed to be carefully preprocessed due to their possible batch effects that may have resulted from the separate performance of DNAm measurements. Distributions of the raw and preprocessed β values by study subsets were visualized for comparison (**Figure 2;Figure 3**). Prior to preprocessing, most of the hypermethylated β values of the KHT set 2 were extremely shifted to the center. As a result of BMIQ normalization and subsequent batch effect correction, the mode of hypermethylated β values of the set 2 became comparable to that of the set 1, suggesting that batch effects between the study sets were successfully addressed. For the downstream statistical analyses, BMIQ-normalized, batch-effect removed DNAm values were used.

In the analysis of differentially methylated CpGs in relation to smoking exposure in the KHT MZ twins ($n=190$), 8 CpGs were significantly associated with smoking at the suggestive significance threshold $p<5e-5$ (**Table 5;Figure 4**). 8 CpGs were annotated to four different loci, *AHRR* (cg05575921, cg23576855, cg21161138), *2q37.1* (in the vicinity of *ALPPL2*; cg21566642, cg01940273, cg05951221), *MYOIG* (cg12803068) and *IL3* (cg04704634). A single CpG within *MYOIG* showed hypermethylation in smokers, while the rest of the top smoking-

associated CpGs were hypomethylated in smokers. Associations at the significance cutoff $p < 0.05$ were observed at 819 CpGs at 687 distinctive genomic loci. 411 sites (50.2%) were hypomethylated in smokers in comparison with non-smokers. However, CpGs with higher significance levels were more likely to be hypomethylated as visualized in the volcano plot (plotting the coefficient against significance levels) in **Figure 4C**. The loci showing most associations with smoking included *AHRR* (11 CpGs), *14q32.33* (7 CpGs), *2q37.1* (6 CpGs), *17q25.3* (5 CpGs) and *1p36.22*, (4 CpGs) and 22 loci, including *CYP1A1*, *GFII*, *HIVEP3*, *MAD1L1* and *MYOIG* (3 CpGs). In the DMR analysis, 5 regions showed significantly differential methylation, including *AHRR* (Chr5:373,299-374,252 and Chr5:368,394-368,447), *2q37.1* (Chr2:233,284,112-233,285,289), *MYOIG* (Chr7:45,002,287-45,002,919) and *KIAA0125* (Chr14:106,329,158-106,329,652) (**Table 8**). Of the 5 DMRs, only *MYOIG* locus showed hypermethylation while the rest of 4 regions were hypomethylated in ever-smokers.

In the analysis of AMDTSS ($n=132$), 5 CpGs were significantly associated with smoking at the suggestive significance threshold $p < 5e-5$ (**Table 6; Figure 5**). 5 CpGs were annotated to three distinct loci, *2q37.1* (in the vicinity of *ALPPL2*; cg21566642, cg01940273, cg05951221), *VARS* (cg17619755) and *AHRR* (cg03604011). Of the smoking-associated 984 CpGs, 671 sites were hypermethylated in the smokers group compared to the non-smokers group. The *AHRR* locus harbored the largest number of smoking-associated CpGs (14 CpGs), followed by *17q25.3* (8 CpGs) and *VARS* (8 CpGs) and *HIVEP3* (7 CpGs). Two

regions showed significantly differential methylation, including *AHRR* (Chr2:233,284,112-233,285,289) and *VAR5* (Chr6:31,760,233-31,760,825) (**Table 8**). The *AHRR* region was hypomethylated with the mean beta fold changes of -0.044 in ever-smokers, while *VAR5* showed hypermethylation with the mean beta fold changes of 0.016.

In the meta-analysis of the KHT and the AMDTSS, 14 CpGs were significantly associated with smoking at the significance threshold of pooled $p < 5e-5$ (**Table 7; Figure 6**). Genomic loci annotated to smoking-associated CpGs include *AHRR* (cg23576855, cg21161138), *2q37.1* (in the vicinity of *ALPPL2*; cg21566642, cg01940273, cg05951221), *MYOIG* (cg12803068, cg22132788) and *F2RL3* (cg03636183). Of the 842 CpGs at 701 distinct loci (pooled $p < 0.05$), 315 sites (37.4%) were hypomethylated (cg23576855 and cg21161138 at *AHRR* and cg01940273 and cg21566642 at *2q37.1*). Meanwhile, CpGs at *MYOIG* (including cg12803068 and cg22132788) and CpGs located at *3p24.3* (including cg03274391 and cg15693572). Of the 842 sites, 14 CpGs were annotated to the *AHRR* loci, 10 CpGs to the *HIVEP3* loci, 7 CpGs to *2q37.1* and 5 CpGs to *17q25.3*.

We compared the EWAS results of the KHT and the AMDTSS for 842 CpGs with pooled $p < 0.05$ of the meta-analysis. Of 842 CpGs, most of the CpGs (840 sites) were consistent in directions of effect sizes between the two sets (**Figure 7A**). In both study populations, hypomethylated CpGs were more likely to be significantly associated with smoking ($p < 5e-5$). CpGs with higher significance levels were more likely to be enriched in KHTS (**Figure 7B**). The *AHRR*

cg05575921 marker (a blue point in the upper right in **Figure 7B**) was highly associated with smoking in KHT ($p=2.1\text{e-}11$), while moderately associated in AMDTSS ($p=6.6\text{e-}5$). However, the pooled p in the meta analysis was 0.002, due to high heterogeneity between the two studies with the effect size of -0.137 (SD: 0.018) for KHT and -0.07 (0.016) for AMDTSS.

For those CpGs differentially methylated, we tested dose-response relationships between DNAm levels and smoking dose-related variables (pack-years, smoking intensity which was defined cigarettes consumed per day and smoking duration (years)). The relationships between the top CpG (cg05575921 for KHT and cg21566642 for AMDTSS) of each of the two studies and smoking dose are presented in **Figure 8** and **Figure 9**. For the KHT population, we observed a steep decrease in DNAm levels of cg05575921 with an increase in pack-years up to approximately 15 pack-years (**Figure 8A**). The response was attenuated among those with high lifetime cumulative dose, after a slight increase of DNAm levels among those exposed to moderate dose. Meanwhile, with increasing smoking intensity up to 20 cigarettes per day, there was monotonous decrease in DNAm levels, followed by a plateau for smoking intensity above ~20 cigarettes (**Figure 8B**). We observed a strong negative dose-response relationship of smoking duration with DNAm levels of cg05575921 up to 20 years of lifetime smoking exposure, followed by moderate positive associations (**Figure 8C**). After cessation of smoking, DNAm levels reverted as a function of years since cessation (**Figure 8D**). For the AMDTSS population, there was a steep decrease in DNAm levels of

cg21566642 with increasing pack-years up to less than ~5 pack-years, followed by positive (~8 pack-years) and negative associations (**Figure 9**).

Finally, we identified traits and ontologies enriched by differentially methylated CpGs using the EWAS knowledge base, EWAS Atlas⁸¹ (**Table 9; Table 10**). A total of 100 traits and 208 ontologies were significant at the converted p -value of modified fisher test < 0.05 . The top traits included smoking, ageing and alcohol consumption. Some of the disease-associated traits/ontologies include cardiovascular risk, lung carcinoma, asthma, multiple sclerosis that were enriched by smoking-enriched CpGs. Environment-associated traits such as exposure to perinatal polychlorinated biphenyls and polychlorinated dibenzofurans, air pollution and arsenic exposure were also enriched.

2.3 Genetic Influences of Smoking-associated DNA Methylation Changes

To examine genetic influences over the methylome variation, we performed DNAm quantitative trait loci (mQTL) analyses for 385 samples of the KHT study whose genotype and DNAm measurements are both available. In the mQTL analysis, 3,609 (19.7%) out of the >18,000 CpGs were under significant genetic influences by a single or multiple SNPs (at Bonferroni-corrected $p < 0.05$) (**Table 11**). One of the 8 CpGs showing associations with smoking (cg04704634; $p = 1.96 \times 10^{-5}$) was under significant genetic control. Meanwhile, 185 (22.1%) out of the 838 CpGs (4 CpGs

excluded due to failure to be mapped to any of the proximal SNPs) that showed significance $p < 0.05$ in this study was under significant genetic control of proximal SNPs (**Figure 10A**). The results of the previous meta-analysis by Joehanes et al. were also examined whether they are under genetic control (**Figure 10B**). 3,609 (19.7%) out of 18,311 CpGs whose proximal SNPs were tested for CpG-mQTL associations were associated with at least one of the mQTLs.

Finally, we tested associations of mQTLs with smoking (ever-smokers with ≥ 10 pack-years). None of the mQTLs achieved genome-wide significance in associations with smoking. The most significantly smoking-associated mQTLs were associated with smoking-unrelated CpGs. (**Figure 11**).

3. Discussion

One of the most consistently replicated exposures that alter DNAm patterns is exposure to cigarette smoke. We have identified DNAm changes induced by smoking exposure using the replication-based approach. Of particular note, we used DNAm measurements of MZ twins, which may have reduced possible effects due to population stratification and genetic polymorphisms^{31, 32}. We further examined whether smoking-associated candidate CpGs were under genetic control.

In this study, we found 842 smoking-associated differentially methylated sites. Among those loci replicated, *AHRR* (Aryl-Hydrocarbon Receptor Repressor) was one of the most differentially methylated, harboring the largest number of significant CpGs. The underlying mechanism of DNAm changes in *AHRR* in response to smoking⁵² is the activation of the arly hydrocarbon receptor (AHR) by smoking-induced PAHs, in which process hypomethylation of *AHRR* is associated with increased expression levels of *AHRR*. In the lung tissue of current smokers, cg05575921 was significantly hypomethylated compared to non-smokers, showing positive correlations in mRNA expression levels of *AHRR*, which was further validated in a mouse model⁸⁴. DNAm at cg05575921 in *AHRR* was also hypomethylated in lung cancer patients in the recent four prospective cohort study²³. Another major locus, *2q37.1*, was located near a cluster of alkaline phosphatase genes, *ALPPL2* (alkaline phosphatase placental-like 2), which are proteins associated with pancreatic carcinoma⁸⁵. *F2RL3* (the coagulation factor II receptor-like 3 gene) which encodes thrombin protease-activated receptor-4 (PAR-4) is

involved in inflammation and pathophysiology of neoplastic and cardiovascular diseases⁸⁶. *MYOIG* (membrane-associated class I myosin) is a gene encoding a protein regulating immune response⁸⁷. Long-coding RNAs (lncRNAs) of *KIAA0125* was reported to be upregulated in ameloblastomas⁸⁸. Such epigenetic alterations across multiple loci reflect not only exposure of smoking, but also increased risk of smoking-attributable diseases in the later life.

The dose-response analysis between smoking-related dose and DNAm levels revealed that DNAm is dependent on dose and time. Though we identified DNAm changes associated with a high dose of ≥ 10 pack-years, changes of DNAm by small dose were observed. The patterns were more dependent on smoking duration rather than daily consumption of cigarette smoking. The DNAm changes persisted in former smokers. However, we observed reversible patterns after cessation of smoking as in prior studies^{15, 89, 90}, though their DNAm patterns were distinguishable with those of never smokers. Persistence of smoking-associated DNAm changes may explain epigenetic link of long-term health effects conferred by smoking, even decades after cessation of smoking⁹¹, while cessation of smoking may possibly help reduce risk of smoking-related health outcomes⁹².

Of the 842 associations between smoking and DNAm in the meta-analyses of the KHT and AMDTSS, we found overall consistency of differential methylations between the two populations. However, several CpGs showed population-specific patterns. For example, cg05575921 (*AHRR*) showed the strongest association with smoking in the Korean population as do many other populations, while we observed

the relatively weaker associations in the Australian population, resulting in the weak signal in the meta-analysis. cg04704634 (*IL3*) whose hypermethylation was previously reported to be smoking-associated also showed strong population-specific patterns, with significant hypermethylation in the KHT smokers and hypomethylation (if insignificant) in the AMDTSS smokers. Meanwhile, two of the best-known smoking-associated sites, cg17619755 (*VARS*) and cg03604011 (*AHRR*), were among the most significantly differentially methylated in the AMDTSS, while we observed no significant associations in the KHT.

Among the smoking-associated DNAm changes detected after controlling for possible confounding factors (age, sex, BMI and cell type heterogeneity), we further examined whether they were under genetic control. In the mQTL analysis, we found that ~20% out of the previously reported >18,000 CpGs were associated with at least one SNP. The larger proportion of the replicated CpGs (22.1%) were under significant genetic control. In the subsequent examinations of associations between mQTLs and smoking, we found that the mQTLs of smoking-related CpGs were not associated with smoking. It was consistent with the finding by Gao X, *et al.* (2017)²⁷, in which mQTLs in the vicinity of smoking-related CpGs were not directly associated with active smoking exposure or all-cause mortality. These mQTLs may modify the DNAm changes due to smoking, contributing to inter-individual variations in DNAm patterns and possible susceptibilities or resistance to diseases conferred by smoking-related DNAm changes.

Taken together, the analyses provided 653 differentially methylated sites in relation to smoking that are independent of effects of mQTLs. These epigenetic markers may reflect smoking-specific changes that are not influenced by DNA sequence variation, one of the major contributors to inter-individual epigenetic variations. However, results should be interpreted with caution, given that cigarette smoke is composed of thousands of chemicals. Different toxicants of cigarette smoke may leave extensive exposure signatures in the DNAm landscape, which may result from combinations of DNAm changes specific to certain chemicals. Some of the DNAm changes in response to cigarette smoking share those due to exposure to environmental dioxin⁹³ or PAHs⁹⁴, for example. Nevertheless, active smoking is one of the most prevalent behaviors that inhale toxicants directly into the body, which triggers detectable response in DNAm compared to those unexposed. Furthermore, the lack of overlap of smoking-associated CpGs with CpGs related with other environmental exposure, such as cadmium exposure⁹⁵ adds to plausibility for specificity⁹⁶.

Major strengths of this study include use of the unique study design using MZ twins and further dissection of genetic control on smoking-associated DNAm changes. Still, there are several limitations to this study. One of the major limitations is possible batch effects that may have resulted from the separate performance of DNAm measurements. To reduce technical biases which could possibly introduce spurious associations, we corrected for batch effects using each batch's positional information. As a result, the distributions of two data sets became

comparable, which made it plausible to combine the data sets for subsequent downstream analyses. Even though this process may have reduced the possible confounding effects tremendously, there still may remain residual batch effects that could not have been addressed with the limited batch information. Second, even though MZ twin studies may confer improved statistical power than unrelated individuals⁹⁷, the moderate sample size may have limited the statistical power. We have applied lenient significance thresholds for scanning smoking-associated CpGs, which may have introduced inflation of type I error. For example, one CpG (cg04704634) among the top CpGs with the p -value threshold $<5e-5$ was associated with one *cis*-mQTL, which may fail to detect CpGs that are independent of effects of mQTL. We further meta-analyzed to resolve the limited statistical power; however, population-specific signals may have been attenuated due to their ethnic differences between two cohorts. Moderate sample size posed limitations in assessing accurate dose-response relationships as well, with only few exposed to high-dose of cigarette smoke, making effects of high dose less confident. Further studies with larger population are warranted for evaluation of valid CpG markers.

In conclusion, we identified the previously reported smoking-associated DNAm sites into those smoking-specific and those under genetic control. Many of the DNAm changes that are specifically induced by exposure to smoking were associated with biological effects of exposure to smoking. Smoking-associated DNAm changes exhibited persistent yet reversible patterns that were time and dose-

dependent. It may provide the underlying epigenetic mechanisms by which exposure to smoking may predispose to long-term adverse health outcomes.

Table 3. Characteristics of the study population of the Korean Healthy Twin (KHT) study

Characteristics	KHT		
	Dataset 1: Discovery set 95 pairs of MZ twins (n=190)	Dataset 2: Validation set I non-MZ twins (n=195)	Dataset 3: Validation set II 61 pairs of MZ twins and their 1st degree relatives (n=149)
Age (years)	48.2 (6.7)	53.8 (16.7)	39.5 (8.1)
Female	96 (50.5%)	96 (49.2%)	82 (55%)
BMI (kg/m ²)	24.3 (3.9)	24.4 (3.5)	23 (2.7)
Number of smokers			
Current	58 (30.5%)	41 (21%)	37 (24.8%)
Former	17 (8.9%)	35 (17.9%)	21 (14.1%)
Never	115 (60.5%)	119 (61%)	87 (58.4%)
Number of ever-smokers (Pack-year > 0)	75 (39.5%)	76 (39%)	64 (43%)
Number of ever-smokers (Pack-year ≥ 10)	60 (31.6%)	57 (29.2%)	35 (23.5%)
Packyears of smokers (years)			
Current	15.5 (7.1)	16.7 (11)	15.9 (11.9)
Former	15.7 (7.9)	31.5 (37.7)	12.8 (12)
Smoking intensity (cigarettes) per day			
Current	15.6 (5.5)	16.1 (8.9)	15.9 (9.2)
Former	18.9 (7.6)	19.9 (13)	8.5 (6.9)
Smoking duration (years)			
Current	20.1 (4.4)	20.9 (6.9)	20.7 (5.8)
Former	15.3 (6.6)	27 (15.7)	16.1 (9.6)
Time since cessation of smoking (years)	6.6 (6.4)	11.5 (9)	8.5 (6.9)
MZ twin pairs discordant for smoking	31 pairs	-	13 pairs
DNA methylation measurement platmforms	Illumina's Infinium HumanMethylation450 BeadChip		Illumina's Infinium MethylationEPIC BeadChip

Table 4. Characteristics of the study population of the Australian Mammographic Density Twins and Sisters (AMDTSS) study

Characteristics	AMDTSS	
	Dataset 1: Discovery set 66 MZ twins (n=132)	Dataset 2: Validation set non-MZ twins (n=347)
Age (years)	55.6 (8.4)	56.7 (7.7)
Female	132 (100%)	347 (100%)
BMI (kg/m ²)	26.5 (8.4)	27 (5.8)
Number of smokers		
Current	12 (9.1%)	29 (8.4%)
Former	37 (28%)	110 (31.7%)
Never	83 (62.9%)	208 (59.9%)
Number of ever-smokers (Pack-year > 0)	49 (37.1%)	139 (40%)
Number of ever-smokers (Pack-year ≥ 10)	23 (17.4%)	52 (15%)
Packyears of smokers (years)		
Current	22.6 (15.2)	18.6 (18.3)
Former	11.5 (11.2)	8.8 (11.2)
MZ twin pairs discordant for smoking	23 pairs	-
DNA methylation measurement platmforms	Illumina's Infinium HumanMethylation450 BeadChip	

Table 5. Top smoking-associated CpGs (KHT, $n=190$; pack-year cutoff=10)

Probe ID	CHR	POS	Ever	Never	Differences	Coefficient	p	FDR	Gene	Annotation
cg05575921	5	373378	0.717	0.866	-0.149	-0.137	2.08E-11	3.84E-07	<i>AHRR</i>	Body
cg23576855	5	373299	0.643	0.777	-0.134	-0.118	4.54E-09	4.19E-05	<i>AHRR</i>	Body
cg21566642	2	233284661	0.38	0.468	-0.088	-0.076	6.03E-07	3.39E-03	<i>ALPPL2</i> *	
cg01940273	2	233284934	0.577	0.641	-0.064	-0.056	7.32E-07	3.39E-03	<i>ALPPL2</i> *	
cg12803068	7	45002919	0.869	0.82	0.049	0.071	4.79E-06	1.77E-02	<i>MYO1G</i>	Body
cg05951221	2	233284402	0.333	0.404	-0.071	-0.053	1.02E-05	3.14E-02	<i>ALPPL2</i> *	
cg04704634	5	131396204	0.722	0.72	0.002	0.046	1.95E-05	5.15E-02	<i>IL3</i>	TSS200
cg21161138	5	399360	0.73	0.782	-0.052	-0.05	4.71E-05	1.09E-01	<i>AHRR</i>	Body
cg22894896	17	29886890	0.45	0.481	-0.031	-0.042	1.10E-04	2.15E-01	<i>MIR193A</i>	TSS200
cg03636183	19	17000585	0.672	0.735	-0.063	-0.058	1.16E-04	2.15E-01	<i>F2RL3</i>	Body
cg03991871	5	368447	0.893	0.93	-0.037	-0.032	2.40E-04	3.98E-01	<i>AHRR</i>	Body

*Not located within a gene region; the nearest gene region presented

Table 6. Top smoking-associated CpGs (AMDTSS, $n=132$; pack-year cutoff=10)

Probe ID	CHR	POS	Ever	Never	Differences	Coefficient	p	FDR	Gene	Annotation
cg21566642	2	233284661	0.411	0.506	-0.095	-0.081	7.40E-06	0.136	<i>ALPPL2*</i>	
cg17619755	6	31760629	0.628	0.612	0.016	0.042	2.17E-05	0.155	<i>VAR5</i>	Body
cg01940273	2	233284934	0.613	0.687	-0.074	-0.056	2.52E-05	0.155	<i>ALPPL2*</i>	
cg03604011	5	400201	0.141	0.116	0.025	0.039	4.80E-05	0.181	<i>AHRR</i>	Body
cg05951221	2	233284402	0.394	0.478	-0.084	-0.066	4.90E-05	0.181	<i>ALPPL2*</i>	
cg05575921	5	373378	0.737	0.811	-0.074	-0.07	6.61E-05	0.203	<i>AHRR</i>	Body
cg08688512	15	37394166	0.183	0.171	0.012	0.026	8.45E-05	0.223	<i>MEIS2</i>	TSS1500
cg15410835	8	143125637	0.61	0.619	-0.009	-0.052	1.05E-04	0.242	<i>TSNARE1*</i>	
cg23366234	21	45713704	0.786	0.784	0.002	-0.027	1.24E-04	0.255	<i>AIRE</i>	Body
cg13716409	2	217924469	0.908	0.906	0.002	0.021	1.48E-04	0.273	<i>ALPPL2*</i>	

*Not located within a gene region; the nearest gene region presented

Table 7. Meta-analyses combining EWAS results of KHTS and AMDTSS

Probe.ID	CHR	POS	Gene	Annotation	Coefficient [95% CI]	Pooled p	Pooled FDR
cg23576855	5	373299	<i>AHRR</i>	Body	-0.113 [-0.142,-0.084]	4.35E-14	8.02E-10
cg01940273	2	233284934	<i>ALPPL2</i>		-0.056 [-0.071,-0.041]	3.11E-13	2.87E-09
cg21566642	2	233284661	<i>ALPPL2</i>		-0.078 [-0.099,-0.057]	5.22E-13	3.21E-09
cg05951221	2	233284402	<i>ALPPL2</i>		-0.058 [-0.075,-0.04]	8.73E-11	4.03E-07
cg12803068	7	45002919	<i>MYO1G</i>	Body	0.059 [0.037,0.08]	7.66E-08	2.82E-04
cg03636183	19	17000585	<i>F2RL3</i>	Body	-0.054 [-0.074,-0.034]	1.46E-07	4.49E-04
cg21161138	5	399360	<i>AHRR</i>	Body	-0.045 [-0.062,-0.028]	3.46E-07	9.11E-04
cg11660018	11	86510915	<i>PRSS23</i>	TSS1500	-0.034 [-0.047,-0.02]	1.40E-06	3.23E-03
cg03965496	1	147718157	<i>NBPF8</i>		-0.046 [-0.066,-0.027]	2.29E-06	4.68E-03
cg25189904	1	68299493	<i>GNG12</i>	TSS1500	-0.062 [-0.088,-0.036]	2.54E-06	4.68E-03
cg00566331	1	12218613	<i>TNFRSF1B</i>		-0.04 [-0.057,-0.022]	7.43E-06	1.25E-02
cg03274391	3	22413232	<i>ZNF385D</i>		0.059 [0.032,0.085]	1.26E-05	1.94E-02
cg05396397	1	11908164	<i>NPPA</i>	TSS1500	0.022 [0.012,0.032]	2.73E-05	3.88E-02
cg22132788	7	45002486	<i>MYO1G</i>	Body	0.028 [0.015,0.041]	3.56E-05	4.69E-02
*Not located within a gene region; the nearest gene region presented							

Table 8. Differentially methylated regions (DMRs) in relation to smoking

Study	CHR	Start position	End position	# of CpGs	Minimum FDR	Mean beta fold changes for ever-smokers (pack-year \geq 10 vs. ref)	Gene symbol
KHT	5	373,299	374,252	4	9.11E-25	-0.065	<i>AHRR</i>
	2	233,284,112	233,285,289	5	2.45E-15	-0.042	<i>ALPPL2</i>
	5	368,394	368,447	2	1.97E-04	-0.026	<i>AHRR</i>
	7	45,002,287	45,002,919	4	2.18E-07	0.036	<i>MYO1G</i>
	14	106,329,158	106,329,652	4	2.71E-05	-0.061	<i>KIAA0125</i>
AMDTSS	2	233,284,112	233,285,289	5	1.25E-10	-0.044	<i>AHRR</i>
	6	31,760,233	31,760,825	10	3.59E-07	0.016	<i>VAR5</i>

Table 9. The top traits enriched by differentially methylated sites in relation to smoking (meta-analysis)

Traits	-log10(p)	Count	Percentage
smoking	>309	516	3.62%
aging	296.25	196	0.79%
maternal smoking	209.63	66	2.02%
smoking cessation	166.49	52	8.09%
educational attainment	161.20	30	41.67%
alcohol consumption	124.71	39	6.07%
HIV frailty	119.37	33	14.41%
down syndrome	103.83	100	0.68%
preterm birth	64.02	63	0.59%
systemic lupus erythematosus (SLE)	56.13	52	0.66%
multiple sclerosis	48.81	50	0.62%
cognitive function	41.00	16	4.09%
IgG glycosylation	38.65	7	100.00%
metabolic trait	34.58	8	29.63%
Kabuki syndrome (KS)	32.48	20	1.06%
perinatal polychlorinated biphenyls and polychlorinated dibenzofurans exposure	31.36	7	35.00%
cleft palate	29.66	12	3.58%
obesity	28.13	37	0.45%
atopy	27.51	32	0.52%
ancestry	27.44	38	0.43%
puberty	26.79	12	2.70%
vitamin B12 supplement	26.76	13	2.21%
cardiovascular risk	25.15	5	62.50%
lung carcinoma	25.11	7	12.50%
B Acute Lymphoblastic Leukemia with t(1;19)(q23;p13.3); E2A-PBX1 (TCF3-PBX1)	24.13	58	0.27%
oral squamous cell carcinoma (OSCC)	23.27	45	0.32%
primary Sjögren's Syndrome (pSS)	23.19	21	0.71%
asthma	21.97	43	0.32%
myalgic encephalomyelitis/chronic fatigue syndrome	21.43	29	0.45%
body mass index (BMI)	19.18	16	0.89%
air pollution (NO2)	17.43	23	0.44%
mortality	17.29	20	0.51%
prostate cancer	17.21	27	0.38%
psoriasis	16.59	20	0.48%
estimated glomerular filtration rate (eGFR)	16.34	7	2.88%
papillary thyroid carcinoma	13.80	24	0.40%
type 2 diabetes (T2D)	13.51	21	0.37%
blood protein biomarker levels	13.10	5	2.98%
Alzheimer's disease (AD)	12.12	10	0.77%
B Acute Lymphoblastic Leukemia with t(12;21)(p13.2;q22.1); ETV6-RUNX1	11.89	18	0.38%
homocysteine levels	10.55	3	2.00%
fetal alcohol spectrum disorder (FASD)	9.70	6	1.07%
Crohn's disease (CD)	9.31	9	0.60%
wellbeing	9.20	2	33.33%
follicular thyroid carcinoma	8.95	17	0.30%

Table 10. The top ontologies enriched by differentially methylated CpGs in relation to smoking (meta-analysis)

Ontologies	-log10(p)	Count	Percentage
smoking status measurement (EFO:0006527)	>309	524	3.39%
smoking behavior (EFO:0004318)	>309	529	3.11%
pack-years measurement (EFO:0006526)	198.95	58	3.48%
aging (GO:0007568)	189.60	149	0.71%
cigarettes per day measurement (EFO:0006525)	171.26	28	40.58%
smoking cessation (EFO:0004319)	166.49	52	8.09%
self reported educational attainment (EFO:0004784)	161.20	30	41.67%
immune system disease (EFO:0000540)	143.08	164	0.53%
autoimmune disease (EFO:0005140)	118.12	116	0.66%
HIV infection (EFO:0000764)	116.41	33	13.10%
genetic disorder (EFO:0000508)	107.84	108	0.65%
Down syndrome (EFO:0001064)	103.83	100	0.68%
drinking behavior (EFO:0004315)	92.53	55	1.20%
alcohol drinking (EFO:0004329)	89.99	51	1.33%
viral disease (EFO:0000763)	77.85	33	3.34%
skeletal system disease (EFO:0002461)	76.42	83	0.55%
nervous system disease (EFO:0000618)	70.52	94	0.48%
cotinine measurement (EFO:0007813)	64.24	17	16.83%
premature birth (EFO:0003917)	64.02	63	0.59%
brain disease (EFO:0005774)	59.62	81	0.47%
rheumatic disease (EFO:0005755)	59.49	58	0.61%
systemic lupus erythematosus (EFO:0002690)	56.13	52	0.66%
carcinoma (EFO:0000313)	55.96	129	0.31%
infectious disease (EFO:0005741)	55.10	45	0.92%
eye disease (EFO:0003966)	52.21	41	0.82%
cardiovascular disease (EFO:0000319)	51.25	114	0.30%
multiple sclerosis (EFO:0003885)	47.97	50	0.60%
digestive system disease (EFO:0000405)	46.64	123	0.30%
body weights and measures (EFO:0004324)	46.07	55	0.52%
hematologic disease (EFO:0005803)	44.15	95	0.30%
lymphoid neoplasm (EFO:0001642)	44.15	95	0.30%
neoplasm of immature B and T cells (EFO:0002425)	43.33	94	0.30%
acute lymphoblastic leukemia (EFO:0000220)	43.33	94	0.30%
B-cell acute lymphoblastic leukemia (EFO:0000094)	43.33	94	0.30%
childhood B acute lymphoblastic leukemia (EFO:1001946)	43.33	94	0.30%
body mass index (EFO:0004340)	43.14	53	0.51%
Rare genetic neurological disorder (Orphanet:71859)	42.12	26	1.06%
epithelial neoplasm (EFO:0006858)	41.86	95	0.31%
cognitive function measurement (EFO:0008354)	40.66	16	3.99%
Rare genetic intellectual disability (Orphanet:183757)	40.10	25	1.03%
Rare genetic intellectual disability with developmental anomaly (Orphanet:183763)	40.10	25	1.03%
serum IgG measurement (EFO:0004565)	38.65	7	100.00%
serum IgG glycosylation measurement (EFO:0005193)	38.65	7	100.00%
Rare genetic developmental defect during embryogenesis (Orphanet:183530)	37.24	25	0.89%
metabolic process (GO:0008152)	34.58	8	29.63%

Table 11. Top smoking-associated CpGs (meta-analysis) under genetic control (Bonferroni-corrected $p < 0.05$)

CpGs					mQTLs					
Probe.ID	POS	Coefficient [95% CI]	p	FDR	SNP	POS	Alleles	Coefficient [SE]	Δ POS	p^*
cg03274391	3:22413232	0.059 [0.032, 0.085]	1.26E-05	1.94E-02	rs117251312	3:22414601	T/C	0.059 [0.023]	1,369	2.13E-02
cg08189186	19:38793546	0.035 [0.018, 0.052]	5.73E-05	6.55E-02	rs75349505	19:38793073	C/G	0.026 [0.011]	473	3.85E-02
cg26099045	2:64291800	0.041 [0.02, 0.062]	1.55E-04	1.10E-01	rs329500	2:64319016	A/T	0.05 [0.015]	27,216	5.91E-03
cg25228737	15:82234347	0.029 [0.014, 0.045]	2.26E-04	1.44E-01	rs4778953	15:82244299	A/G	-0.019 [0.007]	9,952	1.40E-02
cg03884592	1:42384474	0.034 [0.016, 0.053]	3.16E-04	1.72E-01	rs4660585	1:42393001	A/G	-0.033 [0.012]	8,527	1.01E-02
cg15693572	3:22412385	0.039 [0.018, 0.061]	3.34E-04	1.76E-01	rs7640987	3:22435451	T/C	0.034 [0.011]	23,066	2.55E-02
cg20435267	5:178288359	0.017 [0.007, 0.026]	4.37E-04	2.12E-01	rs7722977	5:178282306	G/A	0.011 [0.004]	6,053	1.76E-02
cg20724032	3:41460736	0.02 [0.009, 0.032]	7.08E-04	2.66E-01	rs7609847	3:40534844	A/C	-0.025 [0.006]	925,892	6.12E-03
* p -values of mQTLs (Bonferroni corrected for the average number of markers tested within the corresponding window size)										

Figure 1. Pipeline for preprocessing DNA methylation data

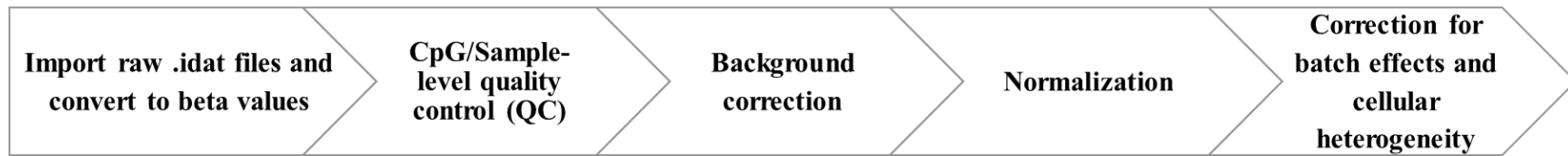
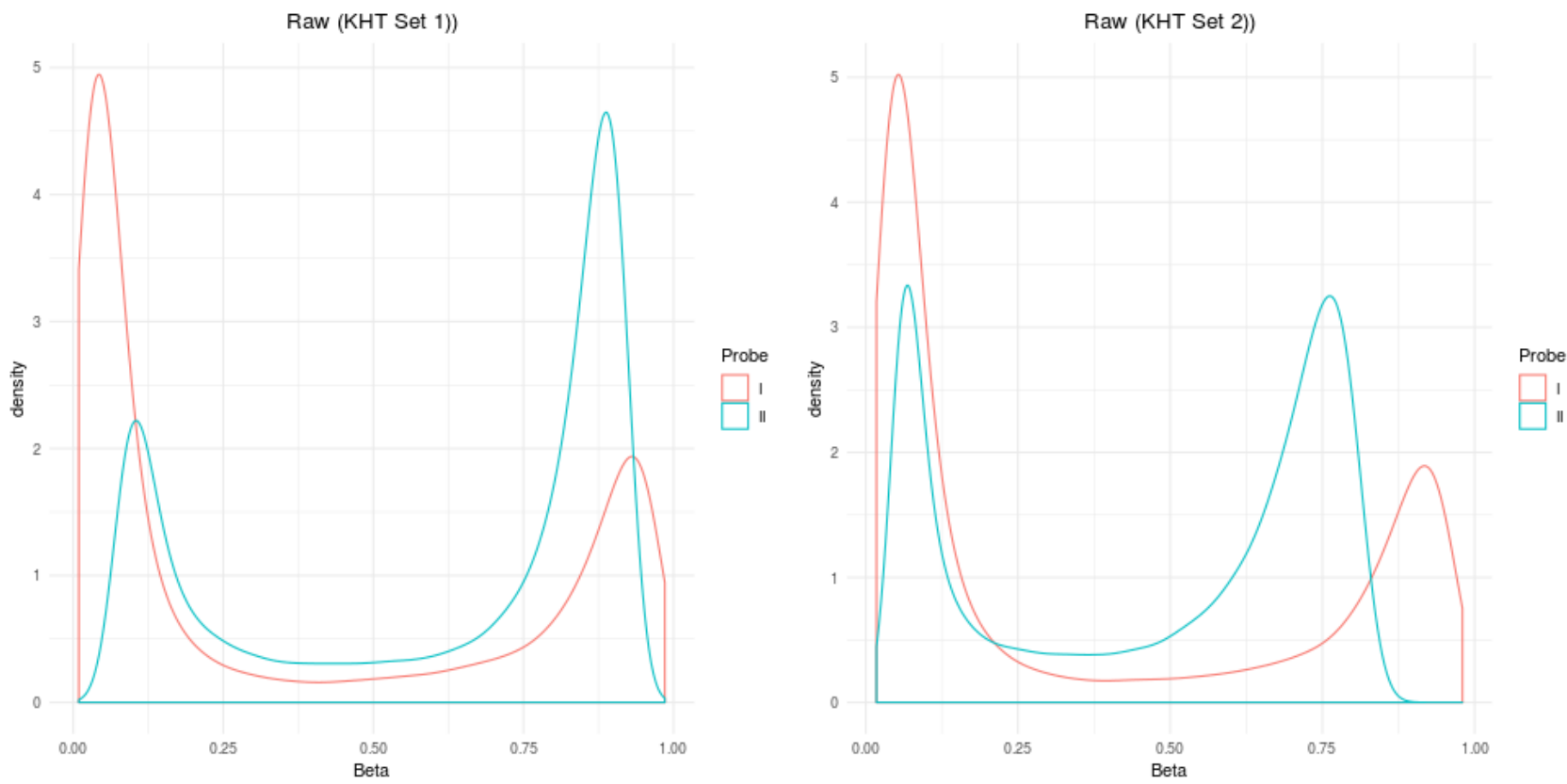
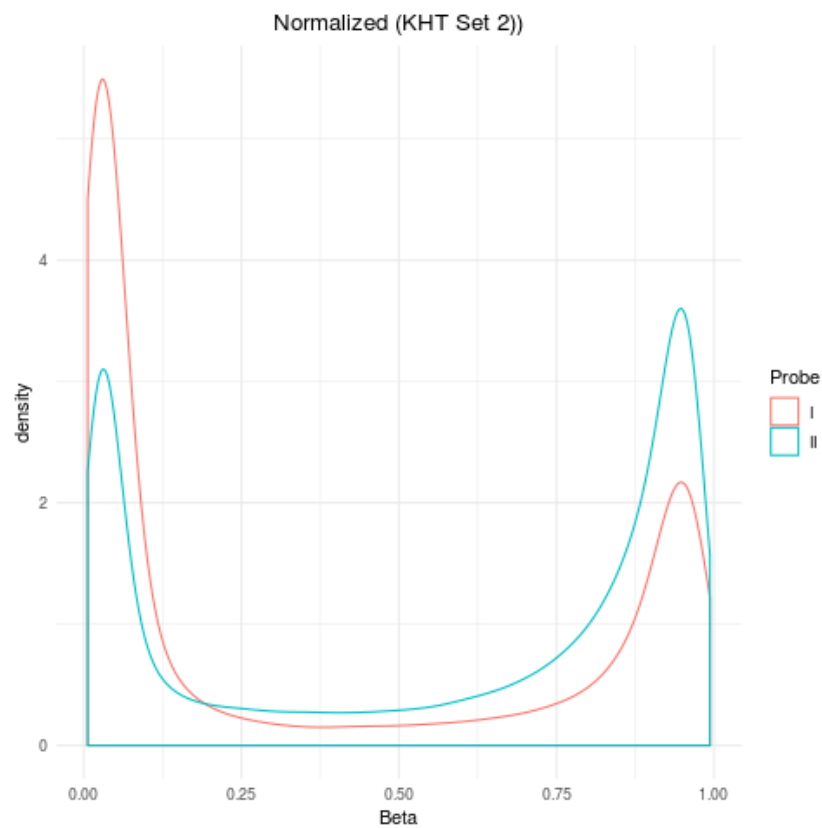
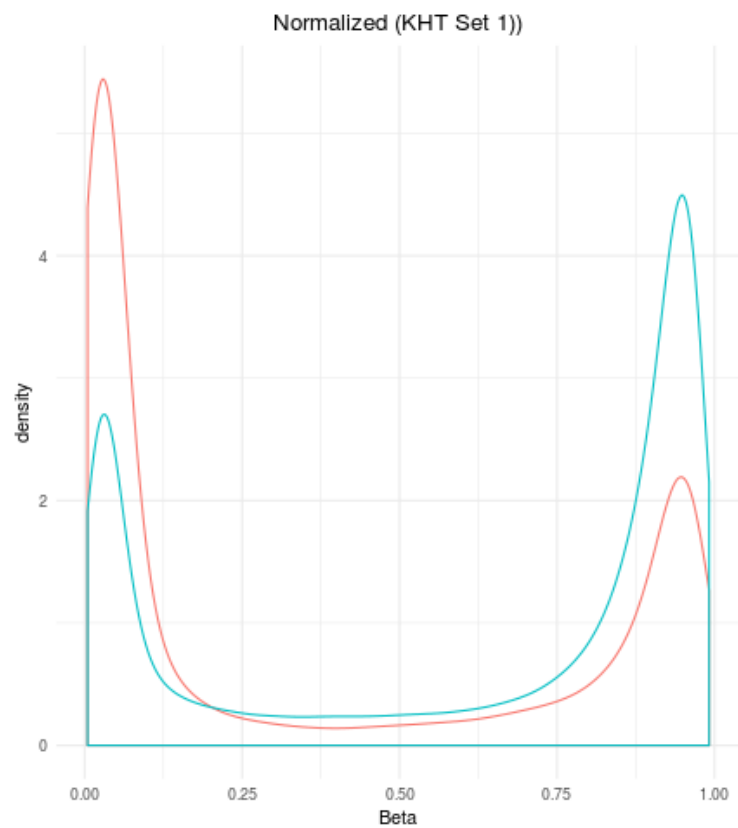


Figure 2. Distributions of all-sample mean DNA methylation values by Infinium probe type (I and II).
A.



B.



* β values before (A) and after (B) BMIQ normalization are presented.

Figure 3. Distributions of all-sample mean DNA methylation values by processing steps (Raw, normalized and batch-effect removed values)

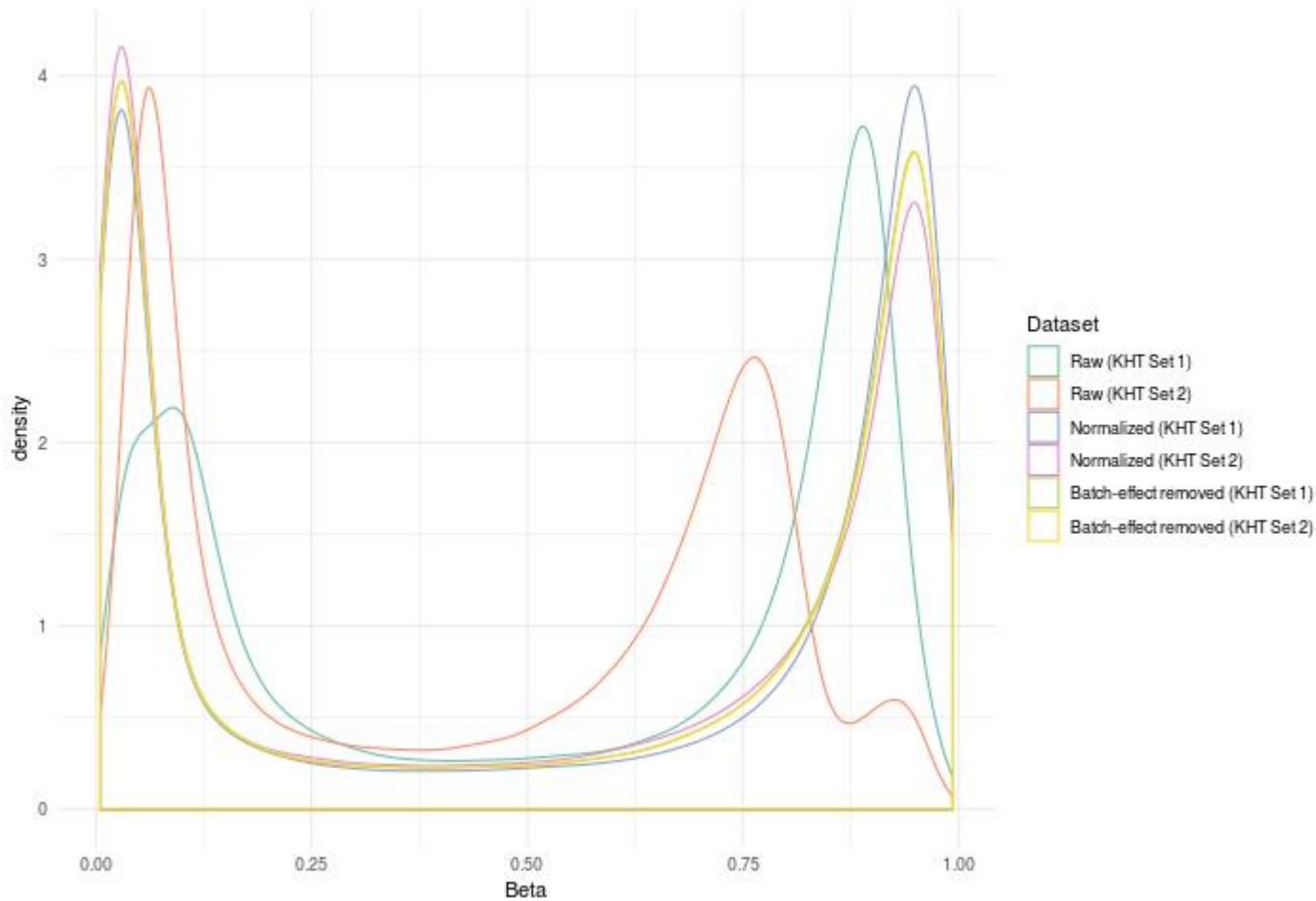
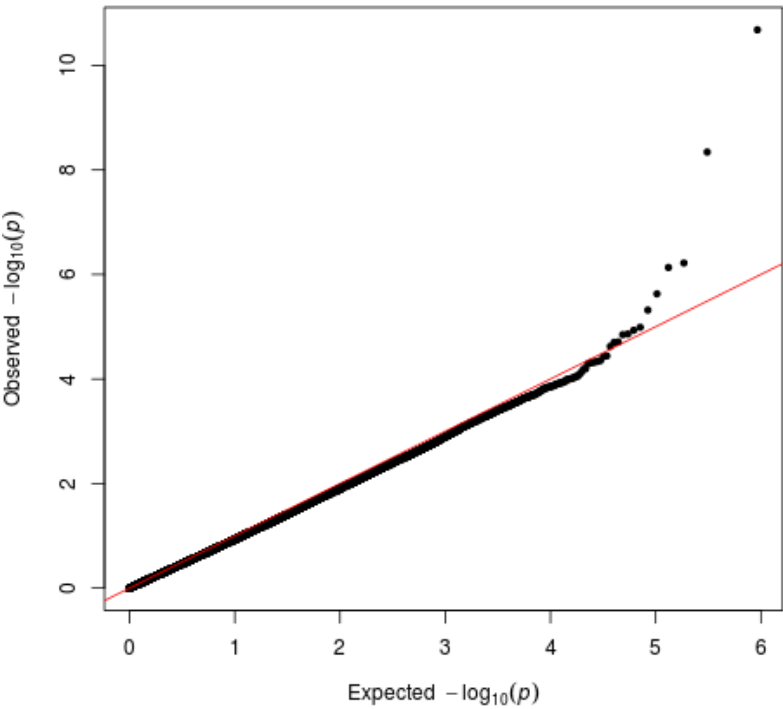
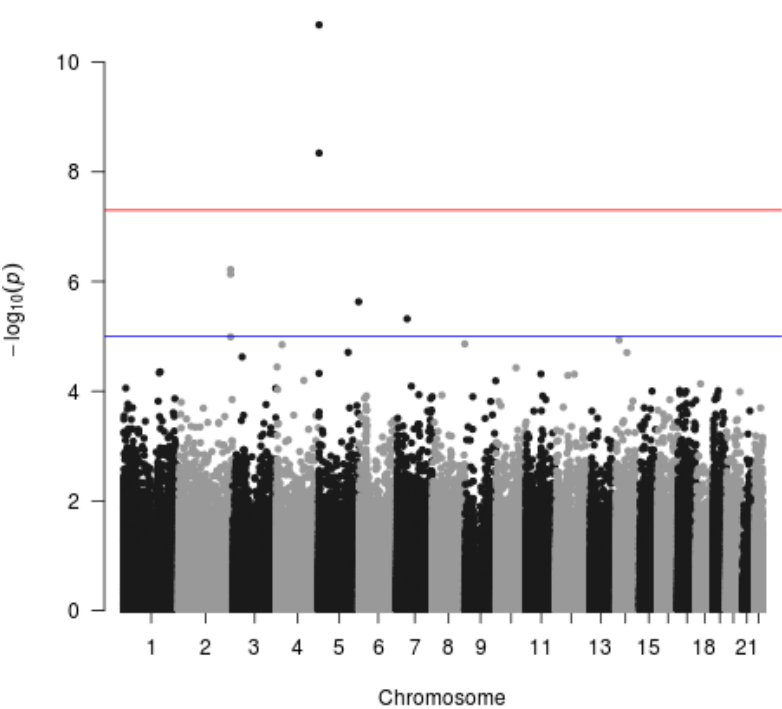


Figure 4. Visualizations of EWAS results of KHT (A. Manhattan plot, B. Q-Q plot, C. Volcano plot)
A.



C.

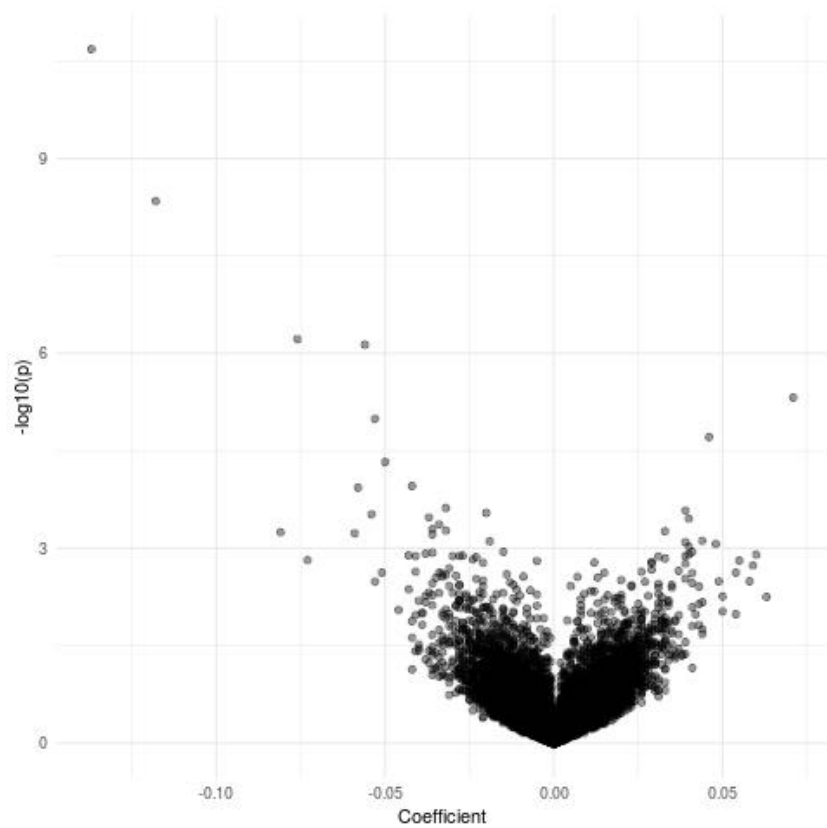
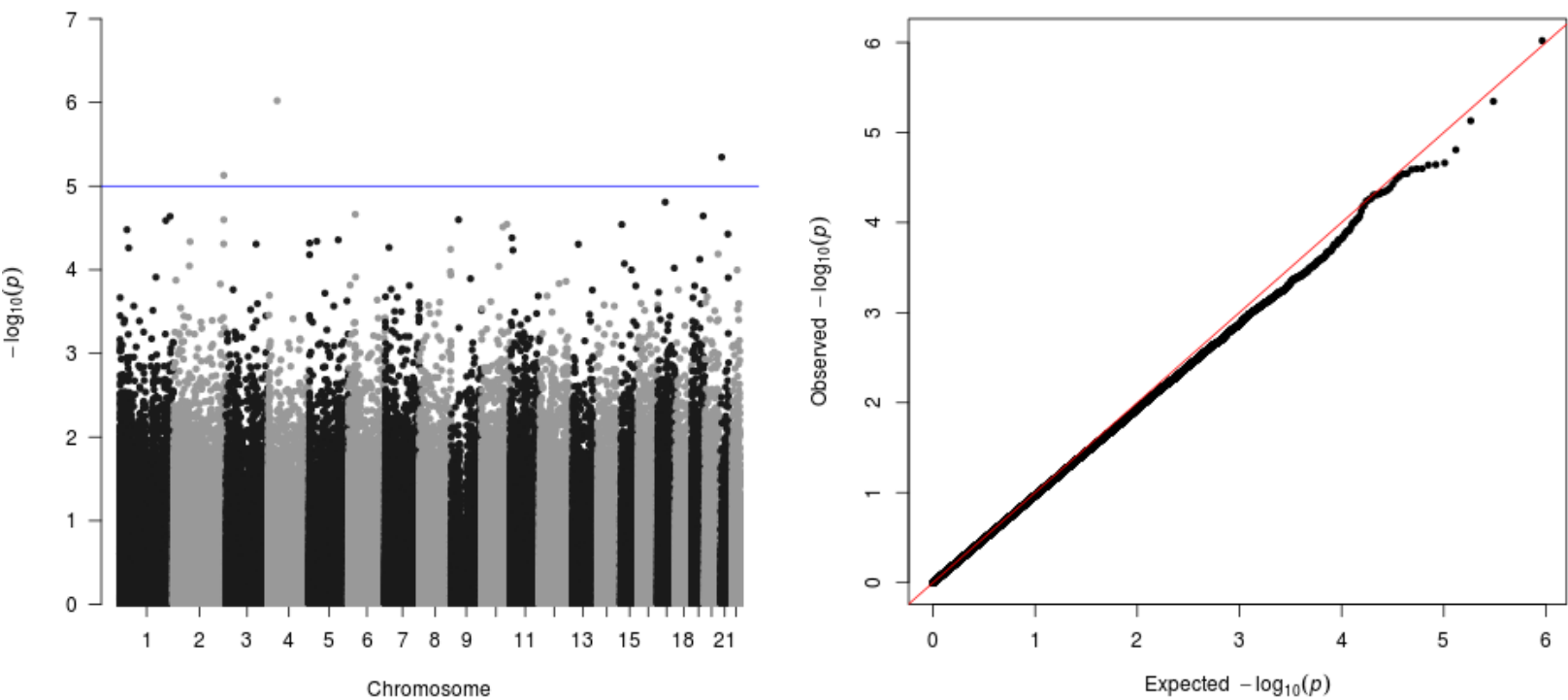


Figure 5. Visualizations of EWAS results of AMDTSS (A. Manhattan plot, B. Q-Q plot, C. Volcano plot)
A. B.



C.

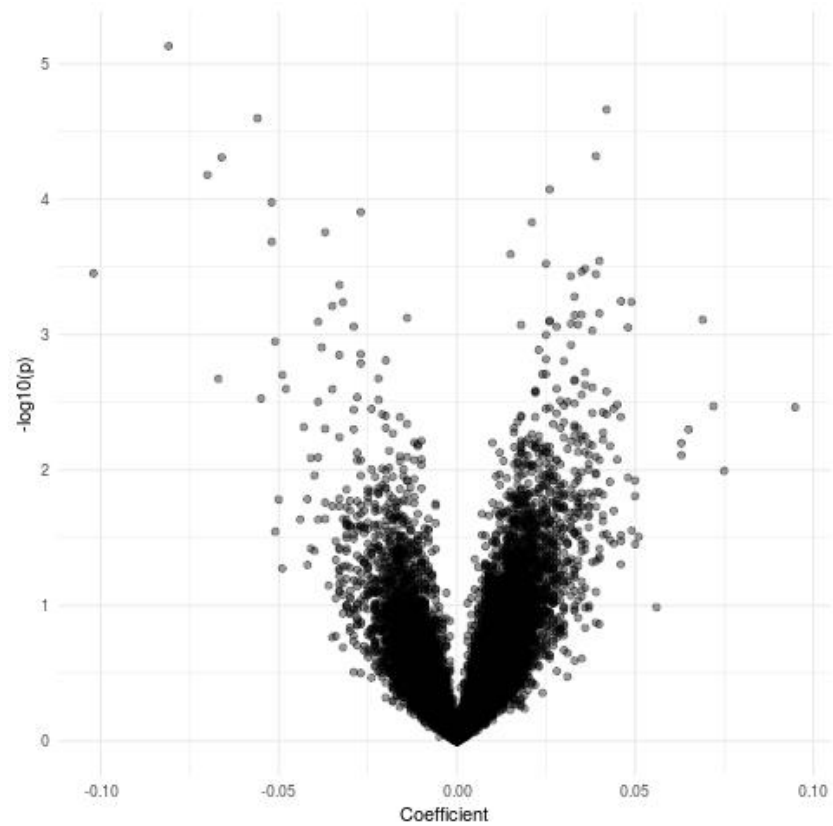
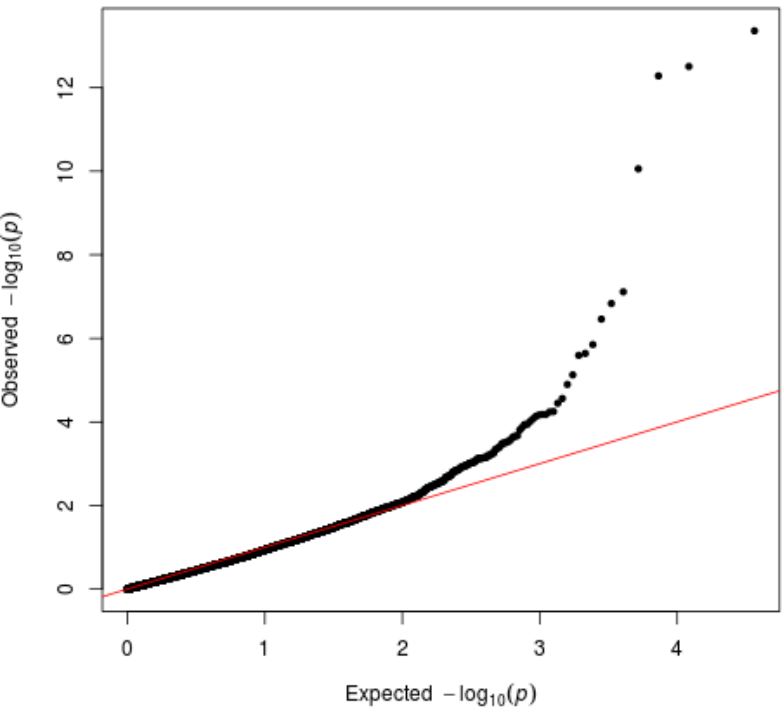
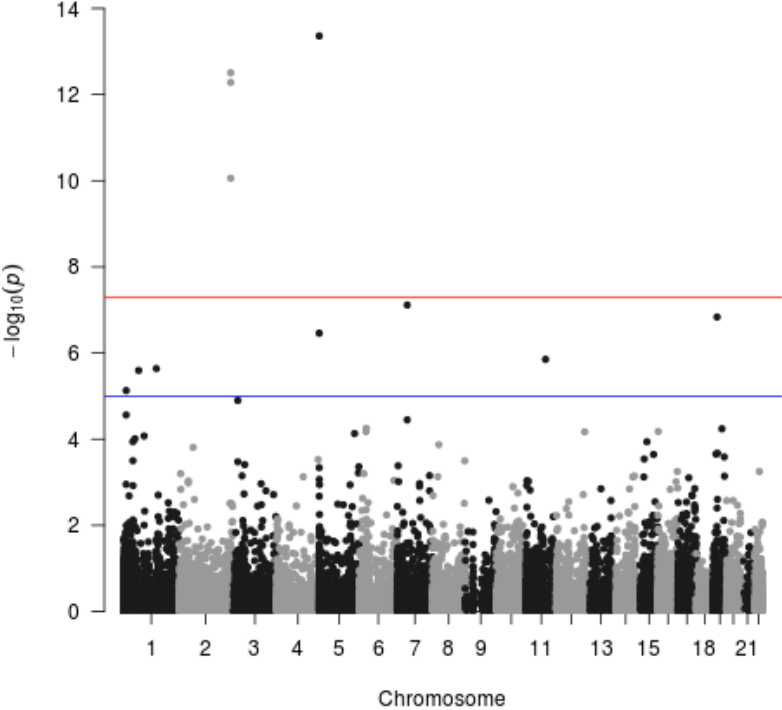


Figure 6. Visualizations of EWAS results of Meta-analysis (A. Manhattan plot, B. Q-Q plot, C. Volcano plot)
A. B.



C.

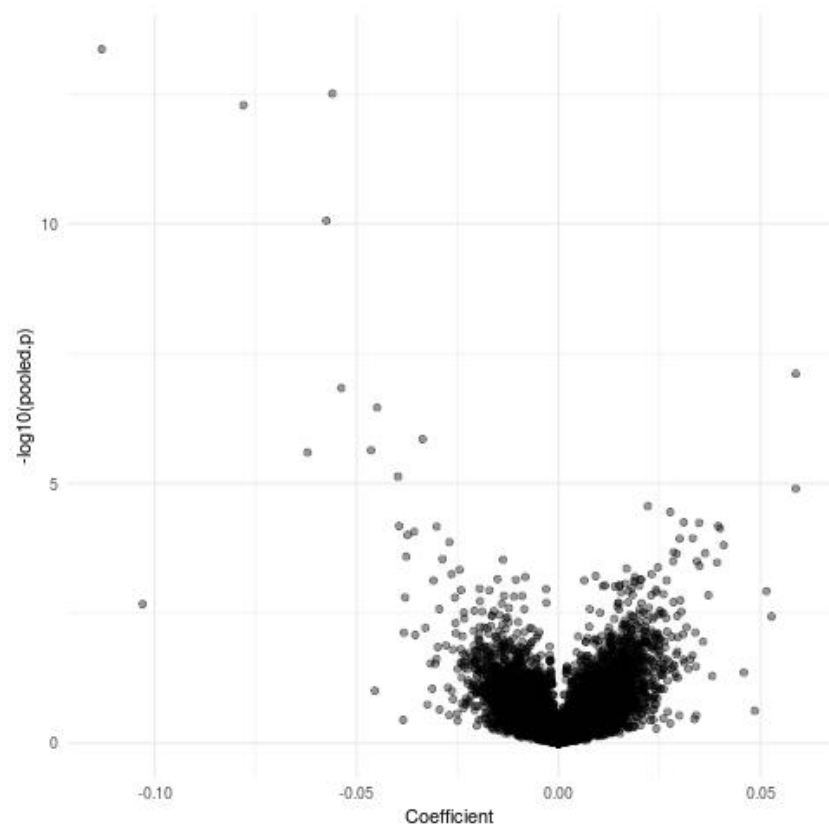


Figure 7. Comparison of effect sizes and significance levels of EWAS results (KHT vs. AMDTSS)

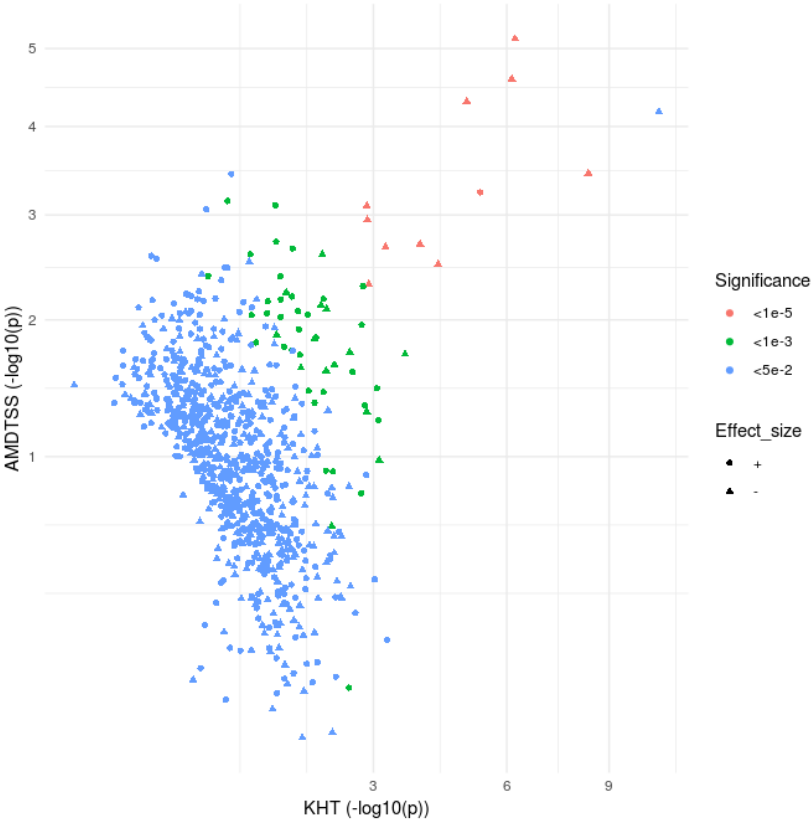
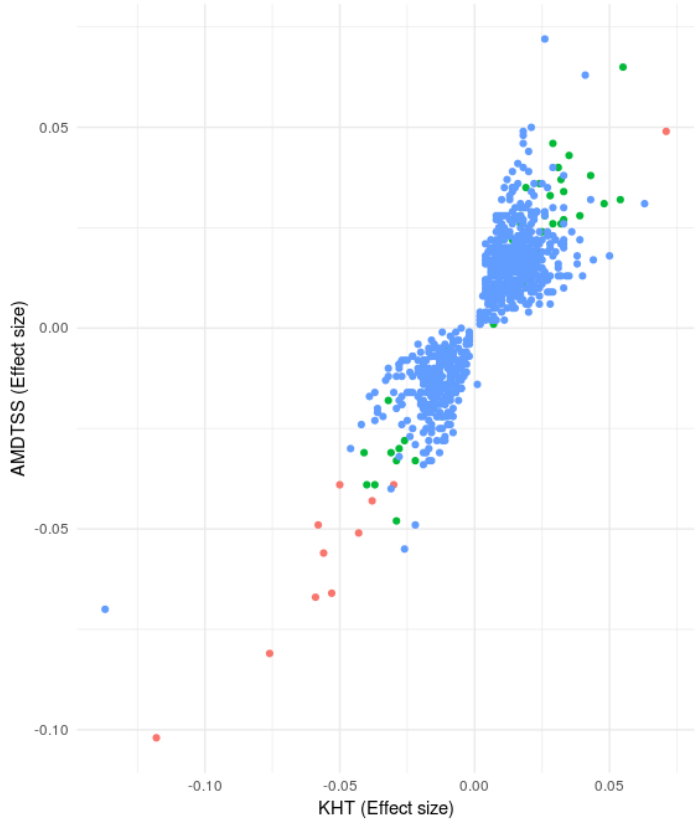
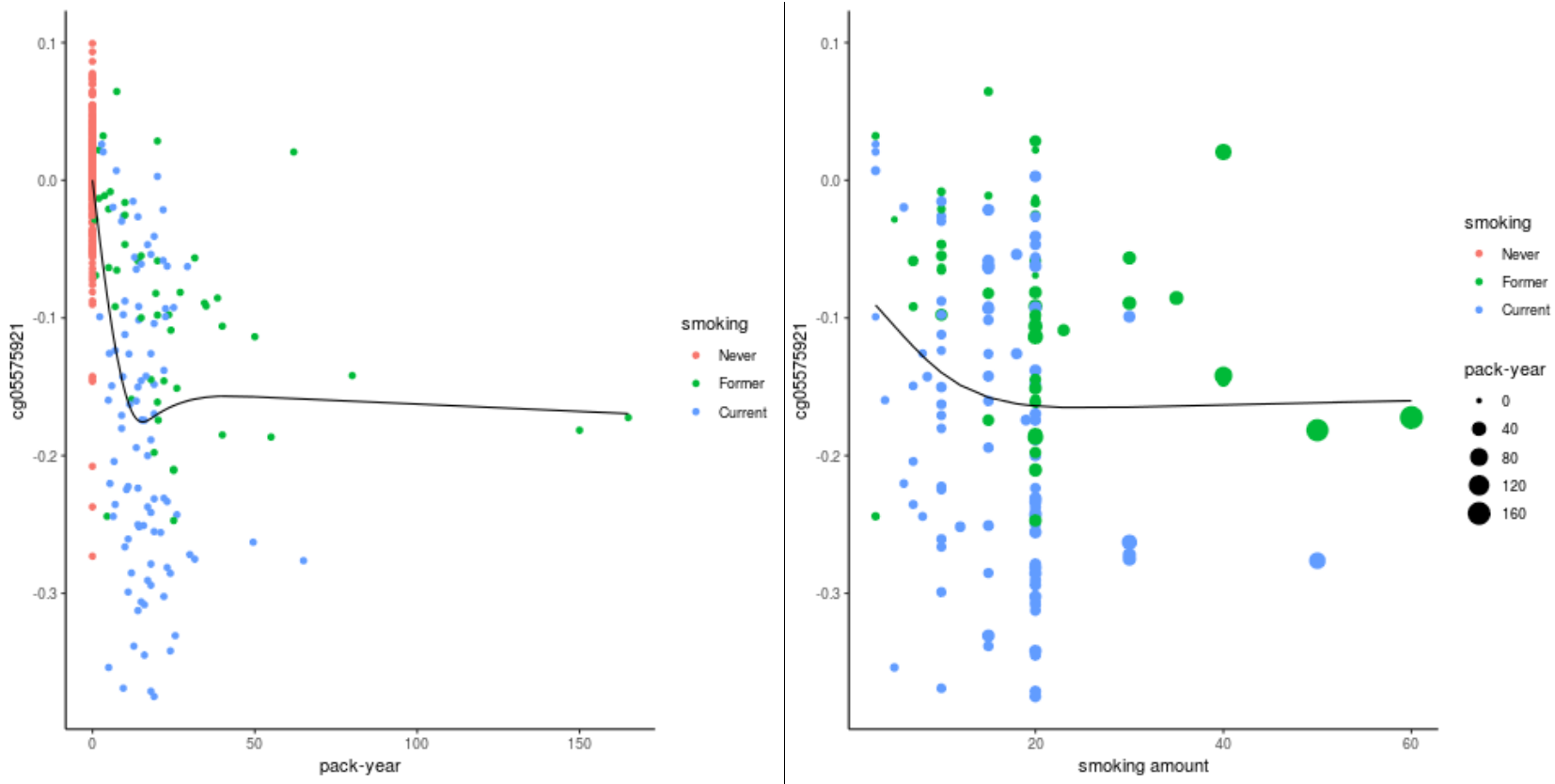
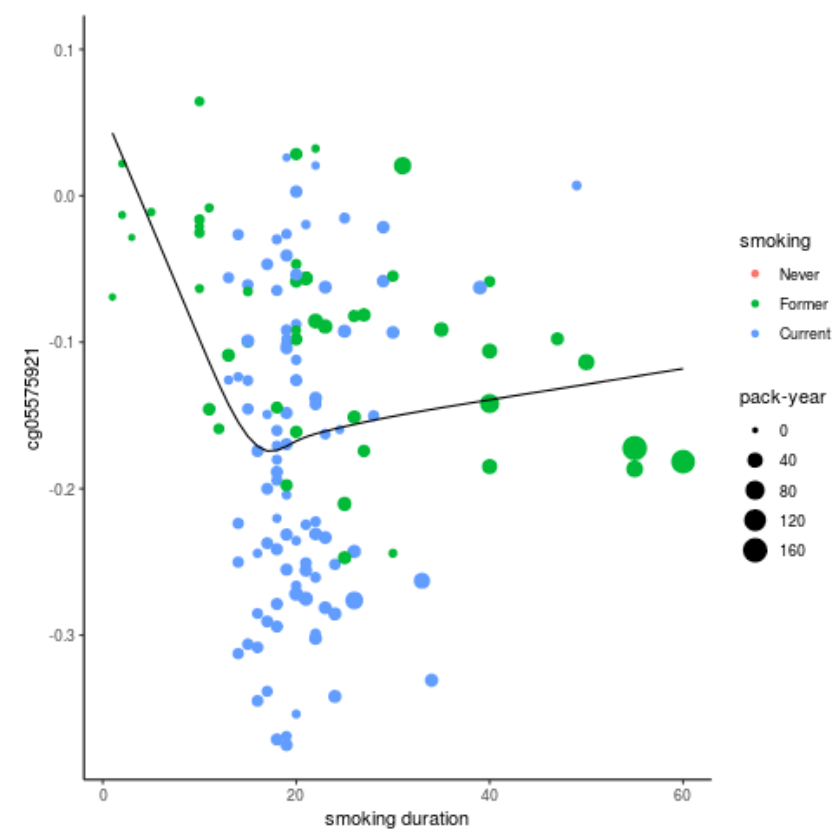


Figure 8. Dose-response analysis of DNAm levels of cg05575921 and smoking-related dose (KHT)
A. pack-years
B. smoking intensity per day



C. smoking duration (years)



D. years since cessation

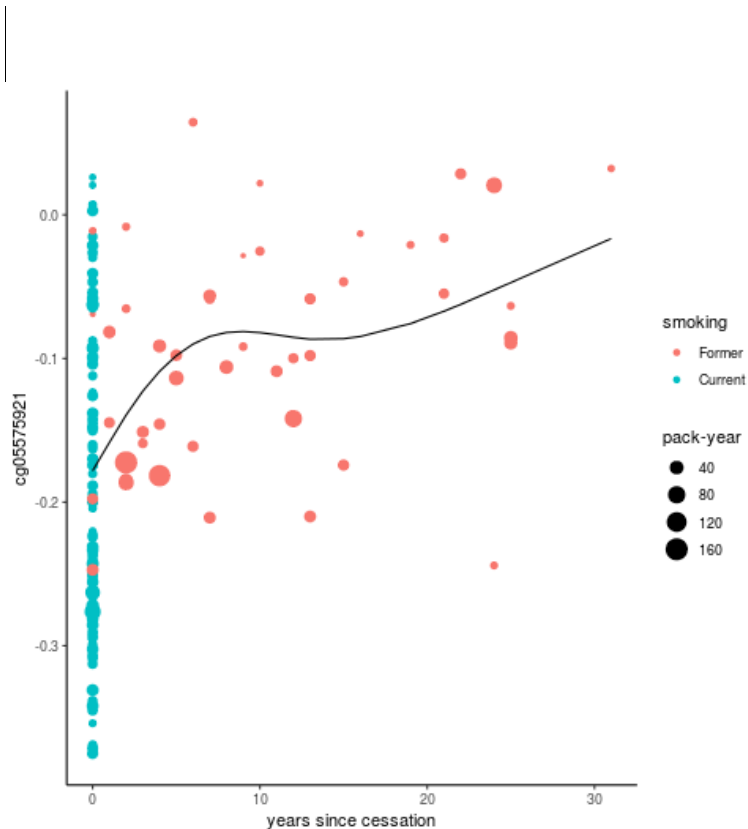


Figure 9. Dose-response analysis of DNAm levels of cg21566642 and pack-years (AMDTSS)

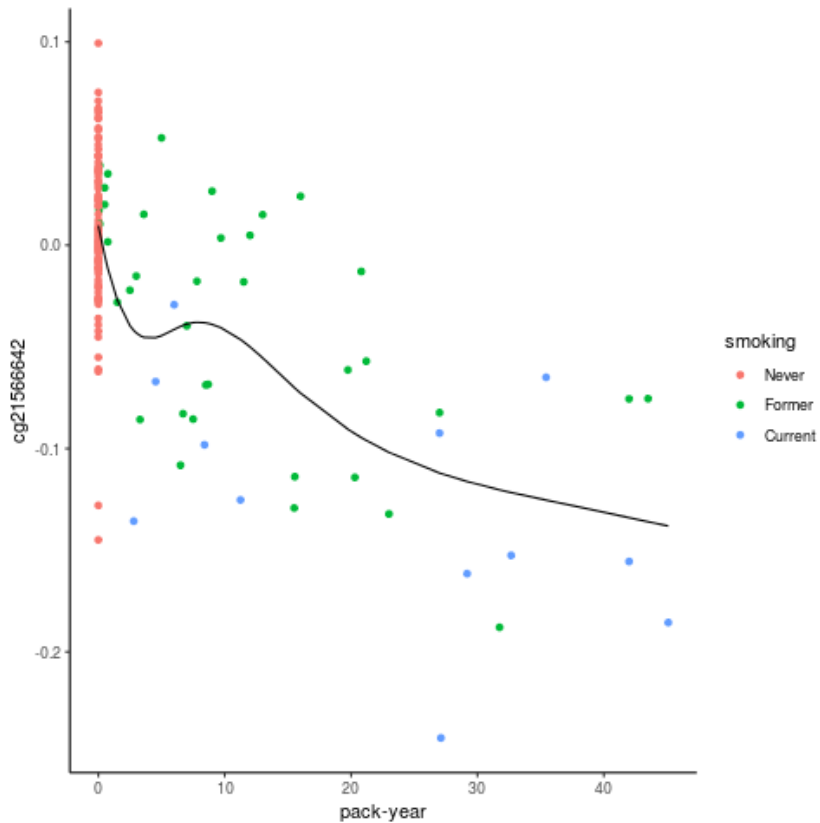
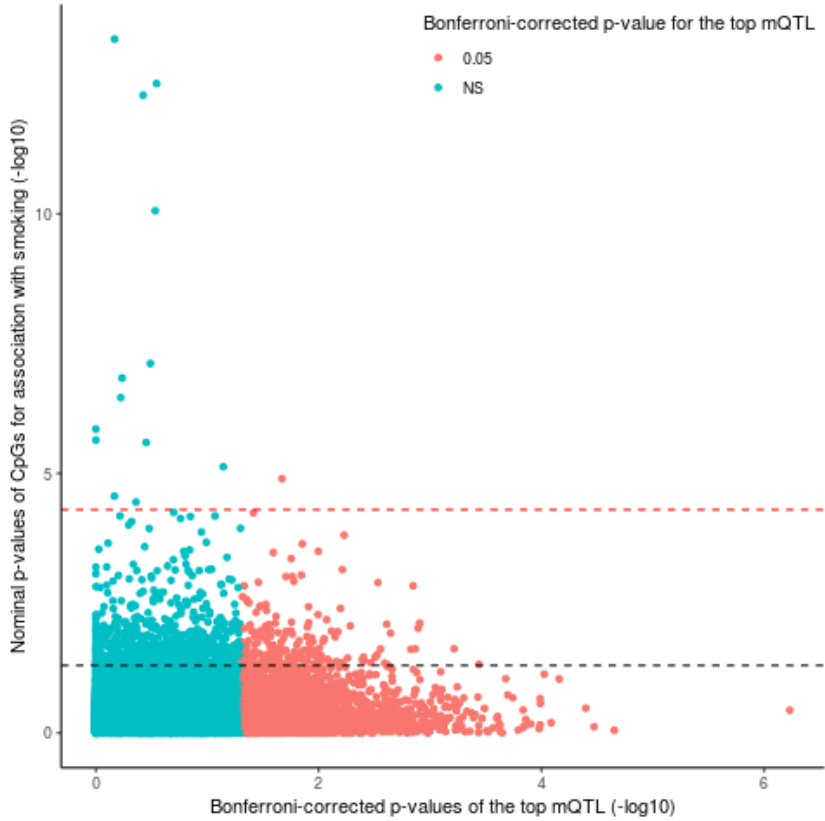
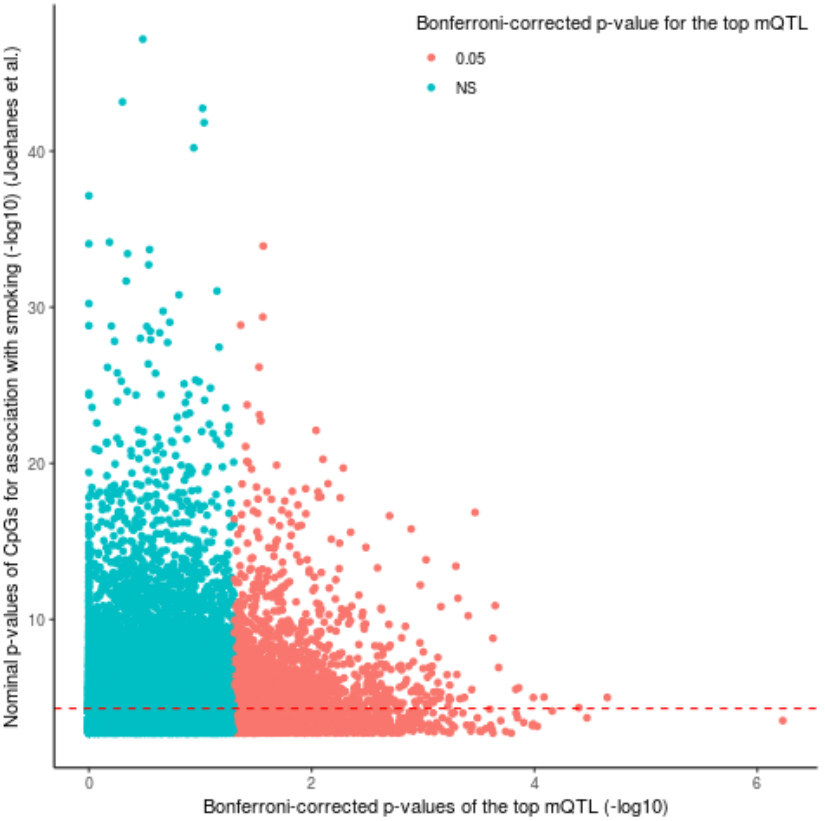


Figure 10. Relationships between significance levels of association between CpG-level DNAm and smoking (x axis) and mQTL significance levels (y axis)
A.

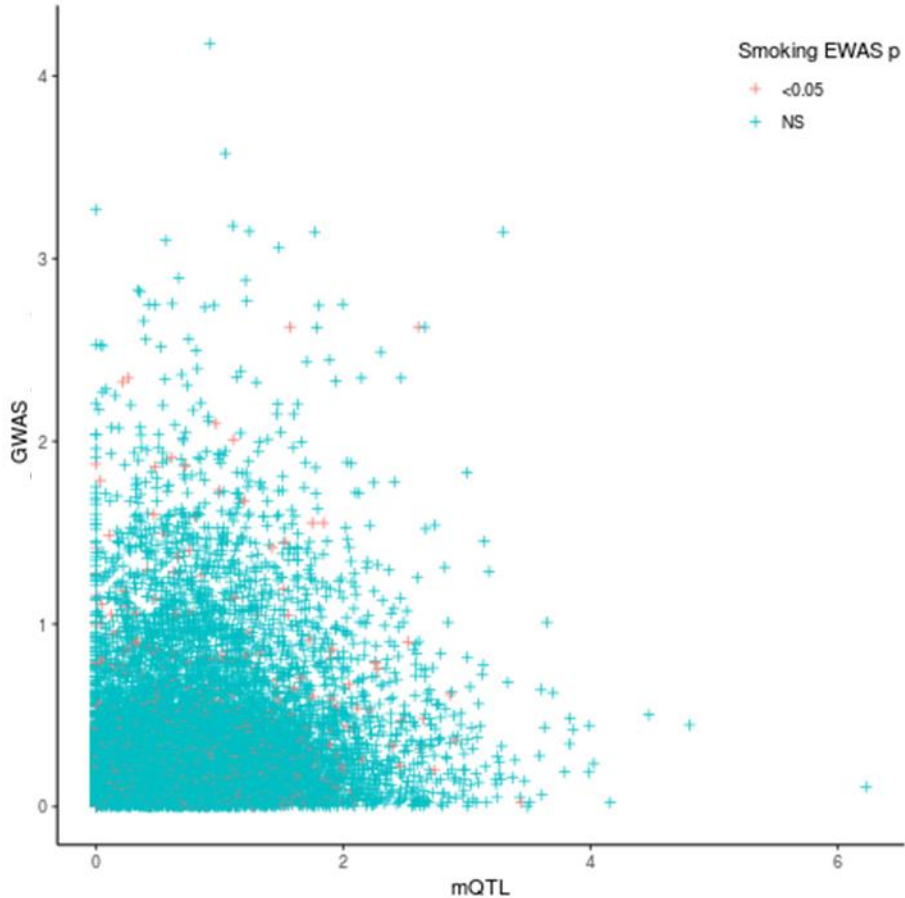


B.



*The red and black lines indicate p -value thresholds of $p < 5e-5$ and $p < 0.05$, respectively.

Figure 11. Associations of mQTLs with smoking exposure



IV. Prediction of Exposure to Smoking Using DNA Methylation Score

1. Material and Methods

1.1 DNA Methylation-based Score

We calculated DNAm score to assess if scoring individuals based on DNAm levels of smoking-associated CpGs is discriminative of exposure to smoking. We utilized the results of associations that included MZ twins of the KHT (training set, $n=190$) (**Table 5**). Particularly, we applied to the validation set I ($n=195$) and II ($n=149$), respectively. The same was applied to the AMDTSS validation set ($n=347$). The characteristics of the samples for the test set are described in **Table 3**.

For the evaluation of DNAm score as a classifier of smoking, we applied the following weighted DNAm score, which is the variation of the method proposed by R Elliott *et al.* (2014)⁹⁸. The basic idea of DNAm-based score is taken from the concept of ‘polygenic risk score’ which was developed to predict risk for polygenic complex traits with a set of multiple SNPs. Polygenic risk score generally uses the number of risk alleles coded as 0, 1 or 2 for bi-allelic SNP variants. The higher number of risk alleles for polygenic risk score can translate into conferring higher risk of diseases of interest. Most of the previous DNAm studies have computed the average β values of the CpG marker of interest or the sum of products of DNAm values and weights without accounting for reference values^{53, 55, 56}. Unlike polygenic risk score based on genotypes, DNAm score makes use of quantitative

measurements of DNAm levels ranging from 0-1, the higher value of which does not translate into higher risk for the trait of interest. To avoid misleading results when computed without the overall DNAm levels for each CpG, we proposed weighted sums of DNAm scores using the deviation from the reference DNAm value. DNAm score (S) for j -th individual for k CpG sites was defined as follows:

$$S_j = \sum_{i=1}^k (\beta_{ij} - (\text{reference } \beta_i)) * \text{weight}_i / k * 100 \quad (\text{Equation 2})$$

where β_i is observed DNAm levels of the j -th individual for the i -th CpG in the test/validation set, reference β is the median DNAm level for i -th CpG and weight_i obtained from the training set is the effect size of i -th CpG on smoking, which can take negative/positive values. The differences between the scores used in this study and the method proposed by R Elliott *et al.* are what reference value was taken and how weighting was defined. While we applied the reference value from the training set, R Elliott *et al.* used the value from the previously reported data. R Elliott *et al.* also used absolute values as weights which they defined as per CpG effect size/average effect size, determining the signs of scores by $(\beta - (\text{reference } \beta))$ for CpGs positively associated with smoking and $((\text{reference } \beta) - \beta)$ for CpGs negatively associated. In this study, we applied effect sizes as weights divided by the number of CpGs included.

Regardless of direction of associations across smoking-associated CpGs, the sum of DNAm score is likely to be higher in smokers, especially when large deviations are observed in CpGs with large effect sizes. For CpGs with positive

effect sizes (weight) (*i.e.*, positive association with smoking), DNAm levels of smoking individuals are likely to be positively deviated from the reference value ($\beta - (\text{reference } \beta) > 0$) multiplied by positive weight, which may result in positive DNAm score. On the other hand, such CpGs with positive weight is more likely to give negative DNAm score to non-smokers, given negative ($\beta - (\text{reference } \beta)$) multiplied by positive weight. For CpGs with negative weight, smokers are likely to have negative deviation and non-smokers positive deviation, both of which are multiplied by negative values as a weight, yielding positive and negative score, respectively.

DNAm score was calculated for CpGs selected according to different significance cutoffs from the EWAS results of the discovery set. First, we calculated score based on the top smoking-associated CpGs with the suggestive thresholds of $p < 5 \times 10^{-5}$. Next, we constructed and compared the following two scoring systems based on the marker set: (i) smoking-associated CpGs without accounting for the effects of mQTLs ($p < 0.05$) and (ii) CpGs with $p < 0.05$ and independent of effects of mQTLs, respectively. We excluded CpGs with any missing values for estimation. All the computational work was performed using the R software (v.3.3.2).

We further assessed dose-response relationships between DNAm score and smoking-related dose (pack-years, smoking intensity which was defined cigarettes consumed per day, smoking duration (years) and years since cessation). Restricted cubic spline regression⁷⁸ was used to identify dose-response relationships between

DNAm score and dose-related variables of smoking exposure. The analysis was performed using the R packages ‘*rms*’⁷⁹ and ‘*Hmisc*’⁸⁰.

1.2 Assessment of Performance of DNAm-based Score

The prediction ability of weighted DNAm score was first evaluated using the area under the curve (AUC) and compared between scores computed using different marker sets. Their classifier performance was visualized using the receiver operating characteristics (ROC) curve⁹⁹. Youden’s Index (J)^{100, 101}, which is defined as (sensitivity + specificity - 1) used for diagnosing accuracy in the prediction models, was acquired to capture the performance. At the maximum J , closest to the top left (‘northwesternmost’) corner of the ROC curve, sensitivity and specificity were measured for comparison. We also additionally compared DNAm-based score with more specific smoking status of current/former/never smokers. Their AUC was also measured using multiclass AUC calculation methods¹⁰². The analyses were performed using the R package ‘*pROC*’¹⁰³.

Improvement in performance by DNAm scores calculated with different marker sets was assessed using three different indices: (i) improvement in AUC^{104, 105}, (ii) net reclassification improvement (NRI)¹⁰⁶ and (iii) integrated discrimination improvement (IDI)¹⁰⁶. Improvement in AUC was evaluated using DeLong’s methods by comparing of AUC of the two ROC curves. NRI examines whether

individuals are classified into higher risk categories (i.e., groups with higher probability of being exposed to smoking) as a result of updating models with different CpG marker sets over the reference. Meanwhile, IDI computes the mean differences in predicted probability of smokers vs. non-smokers over all possible cut-off points between scoring models with different marker sets. The improvement of AUC was calculated using the R package '*pROC*'¹⁰³. NRI and IDI were evaluated using the R package '*PredictABEL*'¹⁰⁷.

2. Results

2.1 DNA Methylation Score by Smoking Status

For the KHT validation set I, we first calculated DNAm score using CpGs with the marker set (i) $p < 5 \times 10^{-5}$ ($n=8$) (**Table 12; Figure 12A**). The mean score was 0.1 in the never-smokers group and 0.848 in the ever-smokers group when the pack-year cut-off was set to be 10, with the difference of 0.747. When calculated with 819 CpGs ($p < 0.05$) (the marker set (ii)), the mean score computed was 0.003 in never-smokers, compared to 0.022 in ever-smokers (difference: 0.005). When excluded mQTL-associated CpGs (the marker set (iii)), DNAm score of ever-smokers was higher by 0.021 (ever vs. never-smokers: 0.025 vs. 0.004). The marker set that excluded mQTL-associated CpGs showed a larger difference between groups (0.021 vs. 0.005). Additional comparison was made between more specific smoking status of current/former/never smokers and DNAm-based risk score. The mean score increased across never, former and current smokers (**Figure 13A**).

We then calculated DNAm score for the KHT validation set II. For the marker set (ii) ($n=7$), the mean score was 0.212 in the never-smokers group and 0.93 in the ever-smokers group, with the difference of 0.718 (**Table 12; Figure 12B**). The mean score for the marker set (i) increased across never, former and current smokers (**Figure 13B**). When calculated with the marker set (ii), the mean score was higher in never-smokers, compared to ever-smokers (-0.044 vs. -0.029). When excluded

mQTL-associated CpGs, DNAm score of ever-smokers was higher by 0.018 (ever vs. never-smokers: -0.034 vs. -0.017).

For the AMDTSS validation set, the mean score computed using the marker set (i) was 0.025 and 0.362 in never- and ever-smokers, respectively (**Table 12; Figure 12C**). The mean score for CpGs ($p < 5e-5$) increased across never, former and current smokers. (**Figure 13C**). The marker set (ii) ($n=737$) yielded the score of -0.044 in never-smokers and -0.029 in ever-smokers. The mean score computed using the marker set (iii) ($n=584$ CpGs) was -0.034 in never-smokers and -0.017 in ever-smokers.

We also examined the dose-response relationships between DNAm score and smoking-related dose (**Figure 14**). For the KHT validation set I (**Figure 14A**), we observed a steep increase in DNAm score with an increase in pack-years up to approximately 15 pack-years. The trend was attenuated among those with high lifetime cumulative dose, after a slight decrease in DNAm levels among those exposed to moderate dose. Meanwhile, with increasing smoking intensity up to 20 cigarettes per day, there was increase in DNAm score, followed by a slight decrease for smoking intensity above ~20 cigarettes. Notably, the larger number who have ever smoked >20 cigarettes per day were those who quit smoking (64% were former smokers). We observed a positive dose-response relationship of smoking duration with DNAm score up to 20 years of lifetime smoking exposure, above which we observed negative and, subsequently, monotonous trends. With increasing years of cessation of smoking up to 10 years, DNAm score was decreased. Among those

who quit smoking 10-15 years ago, there was slight increase in DNAm, followed by decrease for those who quit smoking >15 years ago.

For the KHT validation set II (**Figure 14B**), we observed a steep elevation in DNAm score with an increase in pack-years up to approximately 15 pack-years, after a slight decrease of DNAm score for <5 pack-year ever-smokers. The trend was attenuated among those with high lifetime cumulative dose. Meanwhile, there was increase in DNAm score with increasing smoking intensity up to 15 cigarettes per day, above which little response was observed. With years of smoking, we observed little changes of DNAm for short-term ever-smokers (<10 years), a steep increase for mid-term ever-smokers (10~20 years) and finally a less steep increase for long-term ever smokers (>20 years). Among those who quit smoking, DNAm score decreased up to 10 years of cessation, above which DNAm increased slightly.

For the AMDTSS population (**Figure 14C**), there was overall increase in DNAm score with increasing pack-years up to ~30 pack-years. There was a steep increase for very low-dose smokers with less than <2 pack-years. DNAm score increased less steeply with increasing pack-years. For >~30 pack-year-smokers, little changes of DNAm score were observed.

2.2 Prediction of Smoking Exposure using DNA Methylation Score

The prediction performance of weighted DNAm score was first evaluated using the area under the curve (AUC) and compared between scores computed using different marker sets. For the KHT validation I, we obtained an AUC of 0.917 (CI: 0.878-0.956) (**Figure 15A**). At the maximum J (which is defined as the sum of sensitivity of specificity) of this scoring system, the sensitivity and specificity was achieved at 0.807 and 0.906 (**Table 12**). The cut-off score at the maximum J was shown in **Figure 12A**. The AUC for scoring based on the marker set (ii) was 0.745 (CI: 0.664-0.827), compared with 0.777 (CI: 0.699-0.854) for the CpG set (iii) (**Table 12A**). The AUC for predicting three classes (never, former and current smokers) was 0.849, 0.751 and 0.774 for the marker set (i), (ii) and (iii), respectively (**Table 12**).

For the KHT validation II, we obtained an AUC of 0.895 (0.863-0.953) (**Figure 15B**). At the maximum J of this scoring system, the sensitivity and specificity was achieved at 0.8 and 0.841 (**Table 12; Figure 12B**). The AUC for scoring based on the marker set (ii) was 0.7 (CI: 0.598-0.803), compared with 0.734 (0.632-0.836) for the CpG set (iii) (**Table 12; Figure 15C**). The sensitivity and specificity at the maximum J was 0.788 and 0.834, respectively (**Table 12; Figure 12C**). The AUC when scoring based on the marker set (ii) was 0.61 (CI: 0.524-0.696), compared with 0.646 (0.561-0.732) for the CpG set (iii) (**Table 12; Figure 15C**). The AUC

when predicting never, former and current smokers was 0.85, 0.751 and 0.774 for the marker set (i), (ii) and (iii), respectively (**Table 12**).

In summary, DNAm score based on the top smoking-associated CpGs showed the best performance as a predictor of smoking across all of the validation sets based on different indices including AUC, J statistics, NRI and IDI. Scoring based on the marker set including all markers with $p < 0.05$ without filtering attained moderate performance, while the CpG marker set that excluded mQTL-associated CpGs showed slightly higher performance. At the maximum J , sensitivity was lower than specificity for most of the scores, except for the KHT validation II with CpGs $p < 0.05$. For the KHT validation set I, at the maximum J , both sensitivity and specificity improved after excluding mQTL-associated CpGs. Meanwhile, for the KHT validation set II, we observed that specificity improved, while sensitivity decreased.

2.3 Improvement of Prediction of Smoking Using DNA Methylation Score by Marker Sets

Improvement in performance of predicting smoking based on DNAm scores by different marker sets was first assessed using improvement in AUC (**Table 13;Figure 15**). For all of the validation sets, AUC using the top CpG markers (the marker set (i)) and mQTL-associated CpGs excluded marker set (the marker set (iii)) was significantly higher than the reference marker set (i). When compared

between marker set (i) vs. (ii), the AUC improved statistically significantly by 17.1%, 19.4% and 23% for the KHT validation sets I and II and the AMDTSS validation set, respectively. Meanwhile, the marker set (ii) vs. (iii) yielded the statistically significant yet moderate improvement of 3.1%, 3.4% and 3.7%, respectively.

To measure whether an update of score improves the reclassification ability over the reference (the score based on CpGs <0.05), we additionally estimated NRI and IMI (**Table 13**). For the KHT validation set I, compared to the reference marker set (the marker set (ii)), 8-CpG-based DNAm score gained the statistically significant predictive ability of 45.2% and 29.9% according to NRI and IDI, respectively. The marker set that excluded mQTL-associated CpGs gained the statistically significant improvement of 15.9% in NRI and of 4% in IDI. For the KHT validation set II, 7-CpG-based DNAm score gained the statistically significant predictive ability of 55.9% (NRI) and 28.9% (IDI) compared to the reference marker set (**Table 13**). The marker set that excluded mQTL-associated CpGs gained the statistically significant improvement of 18.1% in NRI and 5.7% in IDI. For the AMDTSS validation set, we observed significant improvement of 41.5% in NRI and 19.5% in IDI when using the marker set (i), while NRI was 5% ($p=0.192$) and IDI was 1.9% ($p<0.001$) for the marker set (iii) comparing to the reference marker set (ii) (**Table 13**).

We further compared AUC values by different p -value thresholds applied to exclude mQTL-associated CpGs from all top N CpGs (**Figure 16**). Overall, we observed decrease of AUC as we include less significant CpGs, across all mQTL

thresholds, except for few scenarios (mQTL p cutoff of 0.2 for the KHT validation set II and mQTL cutoff of 0.5 for the AMDTSS validation set) (**Figure 16**). For the KHT validation set I and II, the best overall performance was observed when CpGs were selected after filtering out CpGs with mQTLs $p < 0.2$. For both KHT sets, CpGs with $p < 0.05$ after excluding mQTLs with $p < 0.2$ achieved the best AUC of 0.812 (KHT I) and 0.782 (KHT II) (**Table 15**). Meanwhile, the AMDTSS set attained the best overall performance when CpGs associated with mQTLs $p < 0.2$ were excluded.

The marker sets whose DNAm score showed the best performance were presented in **Table 14**. For the KHT validation set I, the top 10 to 17 CpGs exhibited superior performance with marginal differences by the number of marker sets, without filtering out mQTL-associated CpGs. With the small set of 3 CpGs, DNAm was excellent in predictions as well. For the KHT validation set II, the same list of the 3 CpGs that excelled in the set I showed the highest performance as well. The following top performance was achieved by mQTL-excluded marker sets. The AMDTSS validation set overall attained high performance by mQTL-excluded marker sets.

3. Discussion

In this study, we have predicted smoking based on DNAm score using multiple validation datasets. The weighted DNAm scoring method proposed in this study is a simple yet informative single predictor, which comprehensively reflects extensive DNAm alterations of each individual across multiple DNAm markers. We demonstrated that the DNAm score was discriminative of smoking status in high-dose ever vs. never/low-dose smokers as well as current, former and never smokers. Particularly, DNAm score based on the top smoking-associated CpGs showed excellent predictive power of smoking. The CpG marker set with lenient association p -value thresholds of 0.05 that excluded mQTL-associated CpGs showed higher performance than the marker set that did not account for effects of mQTLs.

The main focus of this study is on ‘prediction’ of smoking using smoking-related DNAm markers. Those DNAm markers selected for prediction had been identified in the previous chapter, using explanatory modelling, in which association-based statistical models are most commonly applied for hypothesis testing of causal relationships¹⁰⁸. We had thus hypothesized that smoking may induce DNAm changes and tested the hypotheses within the EWAS framework, from which we obtained the list of candidate CpGs to include in the predictive models. In contrast to explanatory modeling, the value of predictive modeling lies in its applied utility, making interpretability of relationships between predictors and outcomes of interest not necessary¹⁰⁸. Instead of detecting well-established DNAm markers of smoking, predictive modeling pursues identification of a set of

informative DNAm markers that predict well. Due to such nature of predictive modeling, a more specific, significant and biologically plausible set of DNAm markers may not likely always lead to increased predictive power. Indeed, the highest predictive power of this study was achieved with a few of the DNAm markers (3 to 20 CpGs), some of which showed better performance without having excluded CpGs under the effects of mQTLs.

Exclusion of mQTL-associated markers contributed to overall significant improvement in discrimination across the validation datasets according to most of the different assessment metrics. AUC, one of the main metrics used in this study, represents summary of ‘overall’ model performance of sensitivity and specificity over all possible thresholds of DNAm score. The cut-off score that best discriminates smokers vs. non-smokers should be carefully selected when using the maximum J , which is a frequently used summary measure of the ROC. At the maximum J of the model that excluded mQTL-associated DNAm markers, specificity was moderate to high ranging 0.75-0.90, at the cost of poor sensitivity of 0.49-0.61, varying by different datasets. In fact, the overall sensitivity of the model that excluded mQTL-associated CpGs was higher than that of the counterpart. Nevertheless, in comparison to a marginal gain in sensitivity, there was a substantial gain in specificity after accounting for mQTLs, resulting in a strong discrepancy between the two measures. The J statistics thus became highly dependent on specificity, as ROC AUC assumes sensitivity and specificity are equally important¹⁰⁹. Although the ROC-based J statistics may enable the selection of an

optimal cutoff value, careful attention needs to be paid when choosing an appropriate cutoff value for prediction.

In order for DNAm score to be utilized as an exposure biomarker of smoking, there are several properties to be examined. They include (1) stability, (2) specificity, (3) dose-response relationships, (4) availability of accessible tissues and (5) availability of measurement technologies. With respect to stability (1), DNAm score reflects persistent changes induced by smoking even decades after cessation. Persistence of DNAm score makes an advantageous biomarker over cotinine, whose half-life is <24h. It was discriminative of former vs. never smokers as well as high-dose ever-smokers vs. never/low-dose smokers. Though specificity (2) can be hardly determined given the complexity of compounds of cigarette smoking, active smoking is one of the most prevalent behaviors that inhale toxicants directly into the body. Furthermore, the lack of overlap of smoking-associated CpGs with CpGs related with other environmental exposure, such as cadmium exposure⁹⁵ adds to plausibility for specificity⁹⁶. The details of specificity of smoking-related DNAm changes was discussed in **Chapter III**. Dose-response relationships (3) were also established by dose- and time-dependent patterns of DNAm score. DNAm score was also reversible after abstaining from smoking. With regard to availability of accessible tissues (4), though the major tissues affected by smoking are lung tissues, other more readily accessible samples such as peripheral blood serve as a useful surrogate for detecting DNAm changes. Toxicants of cigarette smoke circulate thorough the bloodstream via the alveolar capillary system, altering DNAm patterns

of blood⁸⁴. DNAm measurement technologies (5) were well-established, thanks to recent development of next-generation sequencing and microarray technologies, which facilitated rigorous DNAm studies. If not intended to study biobanked genome-wide DNAm data, researchers may find targeted methods such as bisulfite pyrosequencing²⁸, MethyLight²⁹ and EpiType³⁰ are more cost-effective to achieve the specific goal of assessing smoking exposure.

This study successfully demonstrated multi-markers of DNAm is capable of predicting smoking status. However, several limitations of this study have to be acknowledged. First, we evaluated the performance using the validation sets of individuals genetically related with those in the discovery set, which may have possibly introduced biases in evaluating performance. Despite attempts to use different DNAm measurement platforms (HM450K and EPIC chips), testing external validation sets consisting of completely unrelated individuals may provide assessments that are more reliable. Second, most of the models obtained relatively poor sensitivity compared to specificity. One of the possible explanations may be false self-report of current or former smokers that may have contributed to lowered sensitivity. While assessment of smoking based on self-reports shows high accuracy in many epidemiological studies⁴³⁻⁴⁵, there are several populations that exhibit low agreement rates between self-report and objective measures of smoking, including pregnant women⁴⁸ and adolescents^{49, 50}. Especially, in South Korea, where smoking is perceived negatively particularly among females, there was high rate of false response, resulting in underestimated smoking rate in females⁴⁶. We may thus need

further cross-validation of smoking status using other biomarkers such as cotinine levels, even though it may be limited to assessment of current smokers. Lastly, the scoring method used for this study can be limited when the known control samples are included. As DNAm measurements are highly sensitive to between-experiment batch effects, the deviation from the reference value can vary between batches. Even though we may utilize the reference value from the discovery set that can be representative of the study population, it may be challenging to distinguish effects due to smoking from effects of systematic biases due to batch effects. Therefore, the practical application of this scoring method can be limited when valid control samples are not available.

In conclusion, this study presented potential of DNAm changes induced by smoking as biomarkers for detecting exposure to smoking. To the best of our knowledge, this is the first study that accounted for effects of mQTLs in predicting smoking using DNAm markers. DNAm score based on top smoking-associated CpGs acquired from MZ twins achieved high performance, with the good balance of sensitivity and specificity. DNAm-based biomarkers of smoking may be utilized, for example, for legal decision making in lawsuits against tobacco companies. Furthermore, the framework for detecting DNAm exposure signatures of smoking may be applied to identify exposure history of chemical stimuli whose biomarkers are underdeveloped.

Table 12. Performance of DNAm score for prediction of smoking according to marker inclusion thresholds (KHT)

Dataset	Marker inclusion thresholds of associations	Number of markers included	AUC (Never vs. ever) [CI]	Mean score in never-smokers [SD]	Mean score in ever-smokers [SD]	Differences	AUC (Never, former vs. current)	Maximum Youden's Index (<i>J</i>)*	Sensitivity at maximum <i>J</i>	Specificity at maximum <i>J</i>
KHT validation set I	$p < 5e-5$	8	0.917 [0.878-0.956]	0.1 [0.345]	0.848 [0.416]	0.747	0.849	0.713	0.807	0.906
	$p < 0.05$	819	0.745 [0.664-0.827]	0.003 [0.018]	0.022 [0.023]	0.005	0.751	0.43	0.596	0.833
	$p < 0.05$ (mQTL-associated CpGs excluded)	651	0.777 [0.699-0.854]	0.004 [0.018]	0.025 [0.023]	0.021	0.774	0.476	0.614	0.862
KHT validation set II	$p < 5e-5$	7	0.895 [0.863-0.953]	0.212 [0.349]	0.93 [0.497]	0.718	0.82	0.641	0.8	0.841
	$p < 0.05$	737	0.7 [0.598-0.803]	-0.044 [0.017]	-0.029 [0.022]	0.015	0.64	0.321	0.686	0.636
	$p < 0.05$ (mQTL-associated CpGs excluded)	584	0.734 [0.632-0.836]	-0.034 [0.016]	-0.017 [0.022]	0.018	0.66	0.392	0.486	0.907
AMDTSS validation set	$p < 5e-5$	5	0.84 [0.782-0.9]	0.025 [0.236]	0.362 [0.249]	0.336	0.85	0.622	0.788	0.834
	$p < 0.05$	984	0.61 [0.524-0.696]	0.005 [0.015]	0.011 [0.016]	0.006	0.751	0.203	0.365	0.837
	$p < 0.05$ (mQTL-associated CpGs excluded)	794	0.646 [0.561-0.732]	0.004 [0.014]	0.012 [0.015]	0.008	0.774	0.253	0.5	0.753

*Youden's Index (*J*): sensitivity+specificity-1

Table 13. Improvement of performance of DNAm score as a classifier of smoking according to marker inclusion thresholds (KHT)

Dataset	Marker inclusion thresholds of associations	Number of markers included	Improvement		
			AUC	NRI	IDI
KHT validation set I	$p < 5e-5$	8	+0.171 ($p = 3.36e-6$)	0.452 [95% CI: 0.261-0.643] ($p = 0$)	0.299 [95% CI: 0.226-0.373] ($p = 0$)
	$p < 0.05$	819	0.745 (REF)	REF	REF
	$p < 0.05$ (mQTL-associated CpGs excluded)	651	+0.031 ($p = 0.013$)	0.159 [95% CI: 0.052-0.266] ($p = 3.51e-3$)	0.04 [95% CI: 0.027- 0.053] ($p = 0$)
KHT validation set II	$p < 5e-5$	7	+0.194 ($p = 4.21e-4$)	0.559 [95% CI: 0.322-0.797] ($p = 0$)	0.289 [95% CI: 0.199-0.379] ($p = 0$)
	$p < 0.05$	737	0.700 (REF)	REF	REF
	$p < 0.05$ (mQTL-associated CpGs excluded)	584	+0.034 ($p = 0.007$)	0.181 [95% CI: 0.023-0.339] ($p = 0.025$)	0.057 [95% CI: 0.033-0.081] ($p = 0$)
AMDTSS validation set	$p < 5e-5$	5	+0.23 ($p = 8.67e-8$)	0.415 [95% CI: 0.276-0.555] ($p = 0$)	0.195 [95% CI: 0.14-0.251] ($p = 0$)
	$p < 0.05$	984	0.61 (REF)	REF	REF
	$p < 0.05$ (mQTL-associated CpGs excluded)	794	+0.037 ($p = 0.001$)	0.05 [95% CI: -0.025-0.125] ($p = 0.192$)	0.019 [95% CI: 0.01-0.028] ($p = 6e-5$)

Table 14. A list of marker sets that yielded the highest AUC values in predicting smoking

Dataset	# of CpGs before exclusion (EWAS $p < 0.05$)	mQTL p thresholds to exclude CpGs	# of CpGs after exclusion	# of CpGs excluded	AUC [95% CI]
KHT validation set I	14	0	14	0	0.921 [0.883-0.959]
	11	0	11	0	0.921 [0.883-0.958]
	16	0	16	0	0.921 [0.882-0.959]
	12	0	12	0	0.92 [0.882-0.958]
	13	0	13	0	0.92 [0.882-0.958]
	10	0	10	0	0.919 [0.881-0.957]
	17	0	17	0	0.919 [0.88-0.958]
	15	0	15	0	0.919 [0.879-0.958]
	3	0	3	0	0.918 [0.879-0.957]
KHT validation set II	3	0	3	0	0.902 [0.842-0.962]
	3	0.05	3	0	0.902 [0.842-0.962]
	3	0.1	3	0	0.902 [0.842-0.962]
	16	0	14	2	0.9 [0.844-0.957]
	17	0	14	3	0.9 [0.844-0.957]
	7	0	6	1	0.9 [0.845-0.955]
	19	0	16	3	0.9 [0.843-0.956]
	18	0	15	3	0.899 [0.843-0.956]
	2	0	2	0	0.899 [0.839-0.959]
AMDTSS validation set	17	0.05	12	5	0.847 [0.788-0.906]
	9	0.05	7	2	0.846 [0.789-0.904]
	10	0.05	7	3	0.846 [0.789-0.904]
	9	0.1	7	2	0.846 [0.789-0.904]
	10	0.1	7	3	0.846 [0.789-0.904]
	6	0	6	0	0.845 [0.786-0.905]
	6	0.05	6	0	0.845 [0.786-0.905]
	7	0.05	6	1	0.845 [0.786-0.905]
	8	0.05	6	2	0.845 [0.786-0.905]
	6	0.1	6	0	0.845 [0.786-0.905]

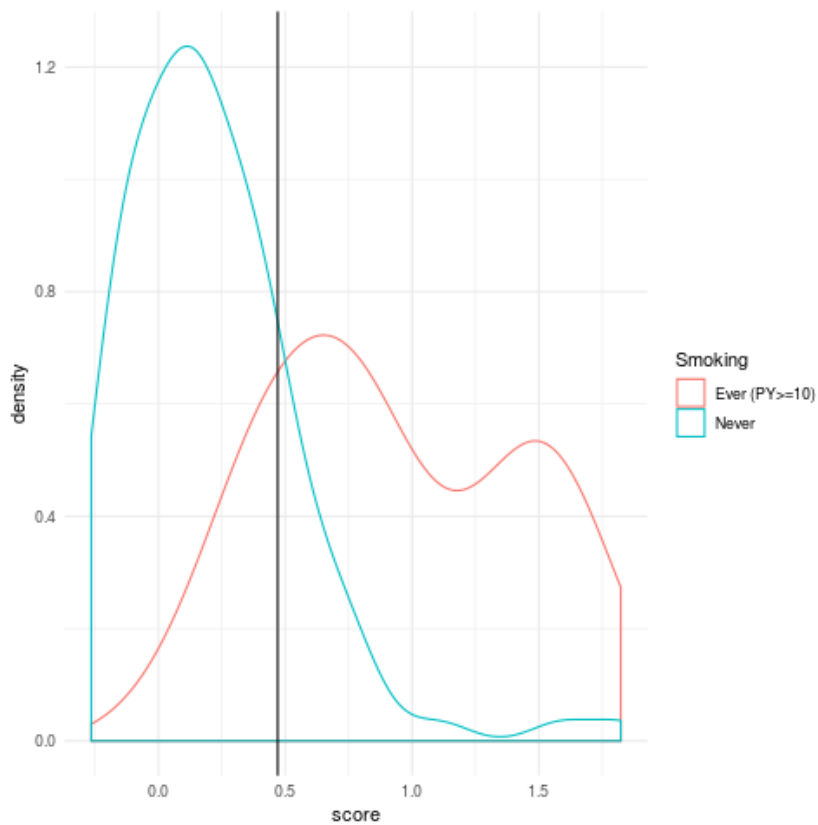
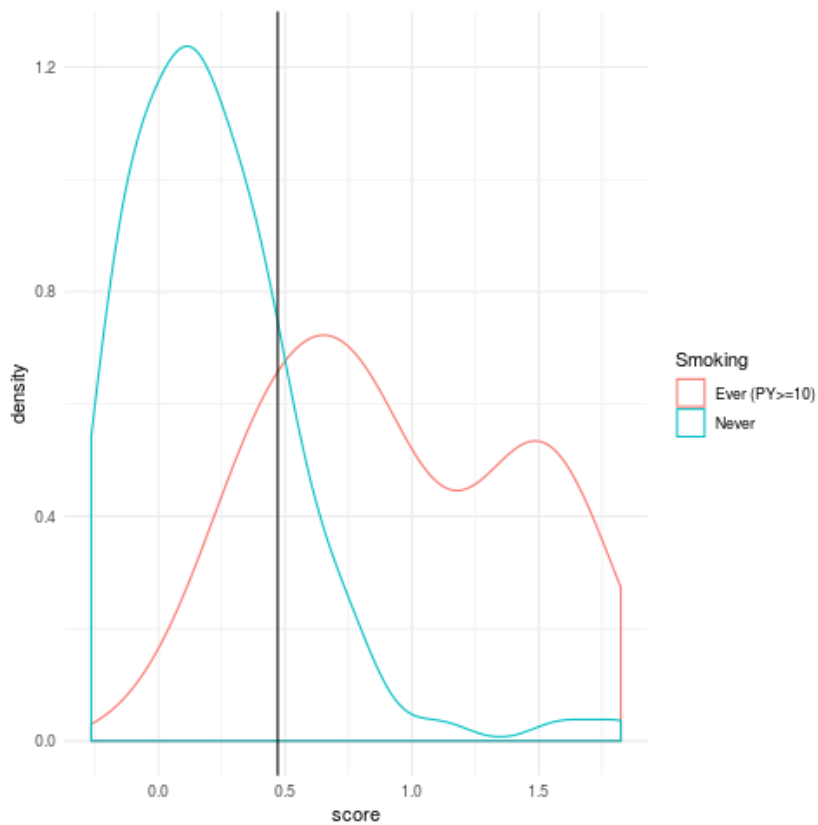
Table 15. Comparison of AUC by different mQTL threshold for exclusion from smoking-associated CpGs ($p < 0.05$)

Dataset	# of CpGs before exclusion (EWAS $p < 0.05$)	mQTL p thresholds to exclude CpGs	# of CpGs after exclusion	AUC [95% CI]
KHT validation set I	819	0	819	0.745 [0.664-0.827]
		0.05	651	0.777 [0.699-0.854]
		0.1	541	0.808 [0.736-0.881]
		0.2	378	0.812 [0.742-0.883]
		0.5	163	0.801 [0.729-0.872]
KHT validation set II	745	0	678	0.711 [0.609-0.812]
		0.05	537	0.751 [0.652-0.849]
		0.1	446	0.778 [0.684-0.873]
		0.2	313	0.782 [0.69-0.875]
		0.5	135	0.738 [0.643-0.834]
AMDTSS validation set	984	0	984	0.61 [0.524-0.696]
		0.05	794	0.646 [0.561-0.732]
		0.1	669	0.649 [0.564-0.734]
		0.2	462	0.683 [0.599-0.767]
		0.5	192	0.669 [0.587-0.752]

A. KHT validation set I

A. KHT validation set I

B. KHT validation set II



C. AMDTSS validation set

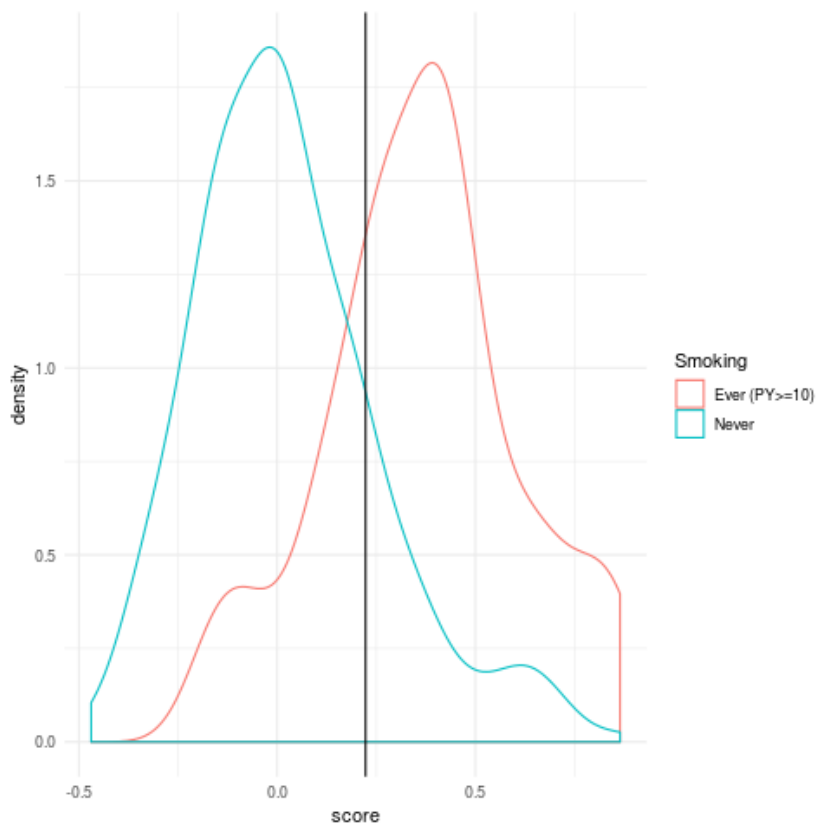
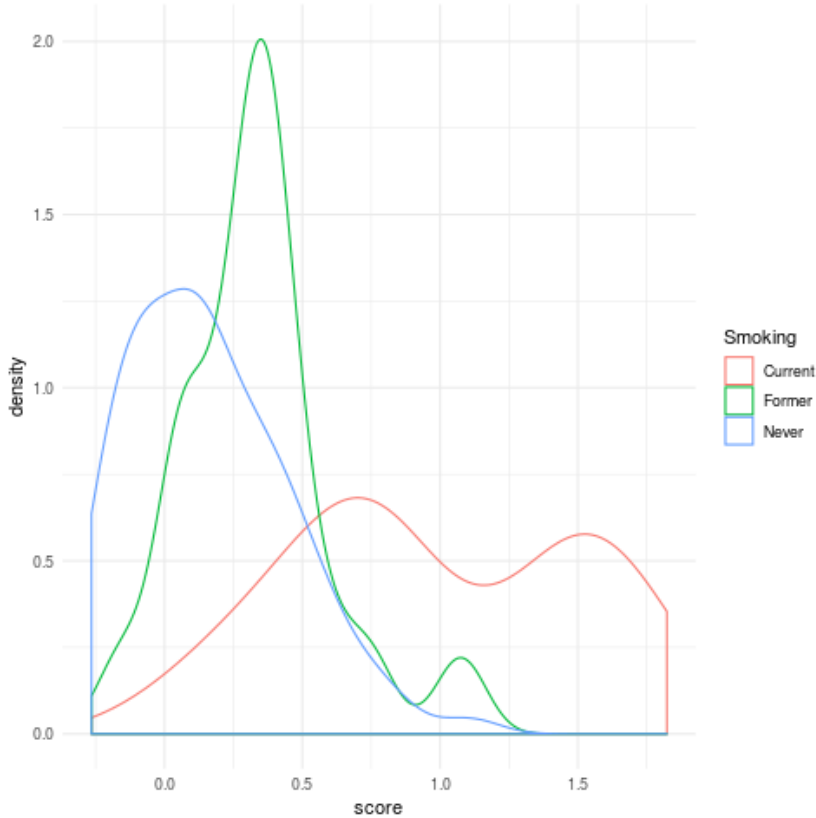
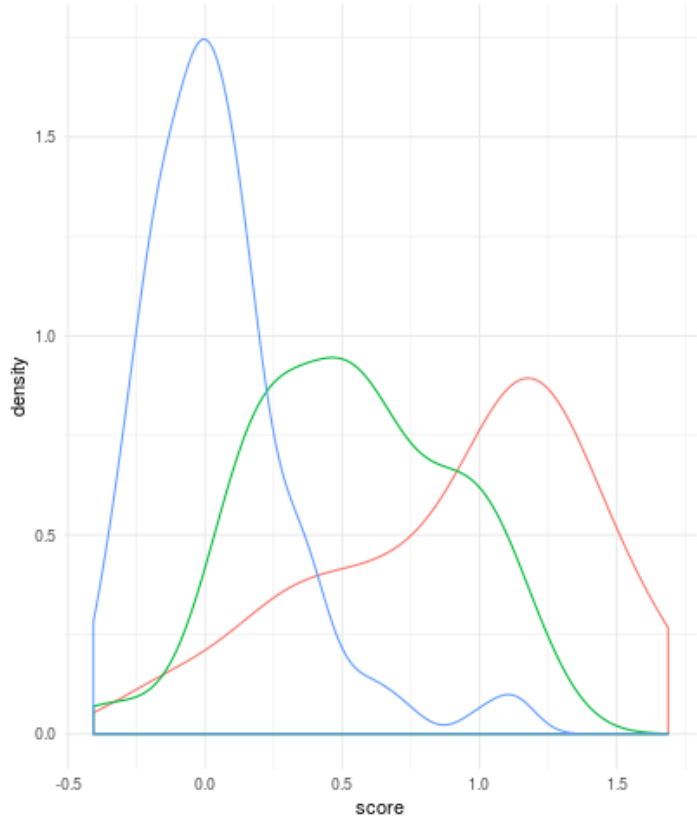


Figure 13. Distributions of smoking-associated DNAm score by smoking status (Current/former/never smokers) ($p<5e-5$)
A. KHT validation set I
B. KHT validation set II



C.

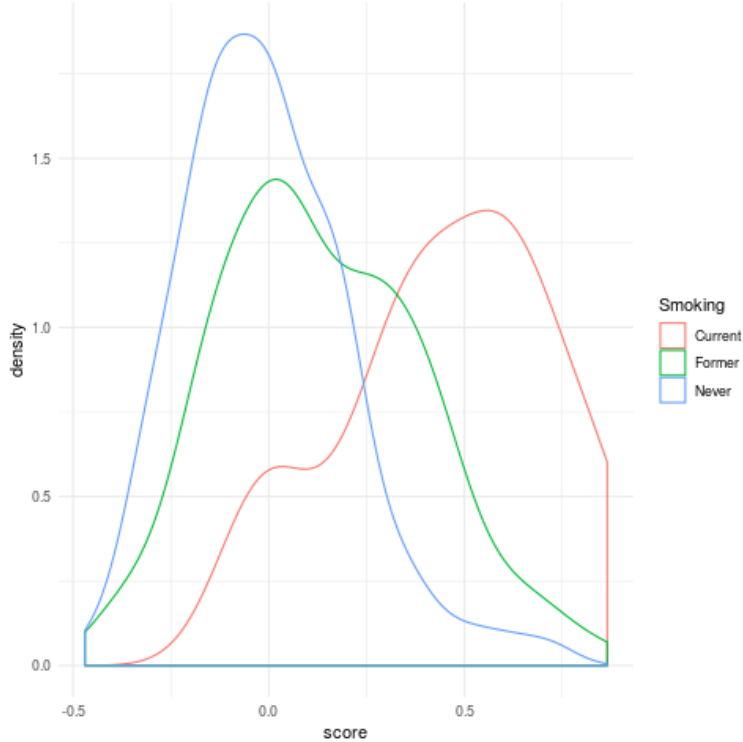
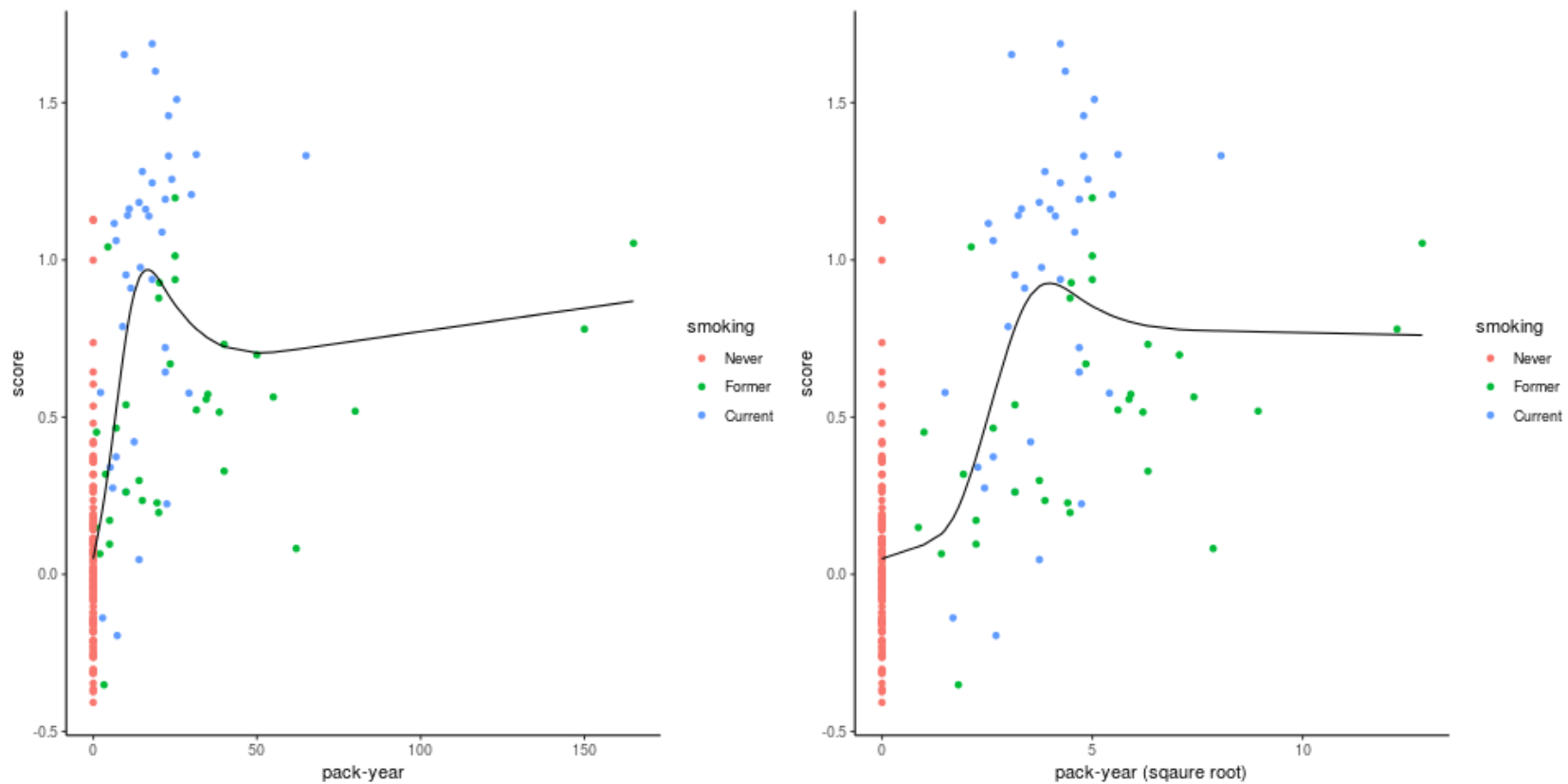
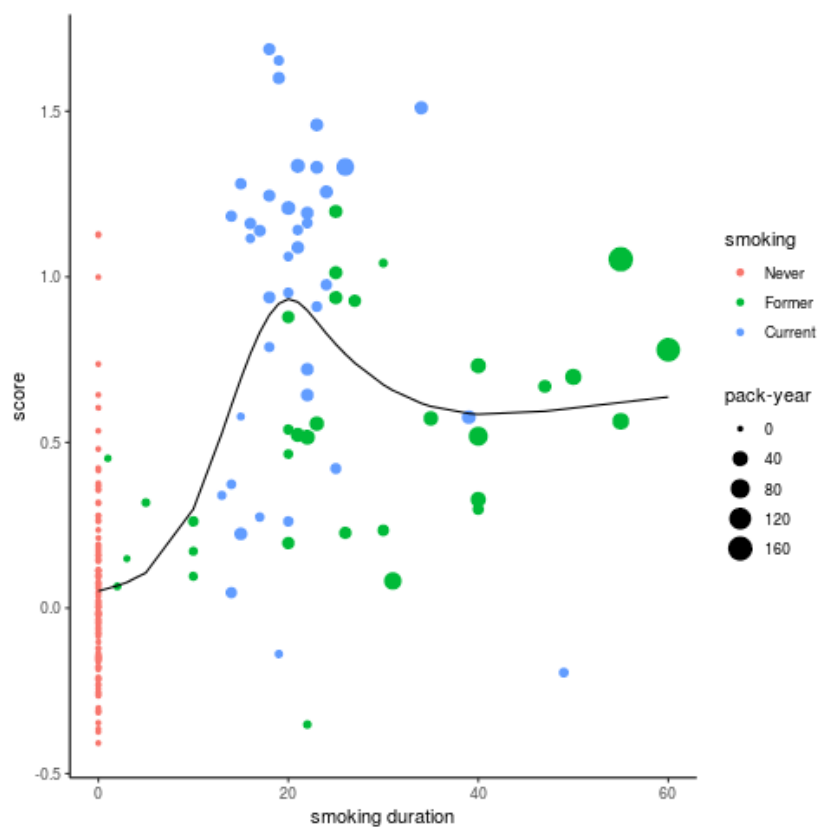
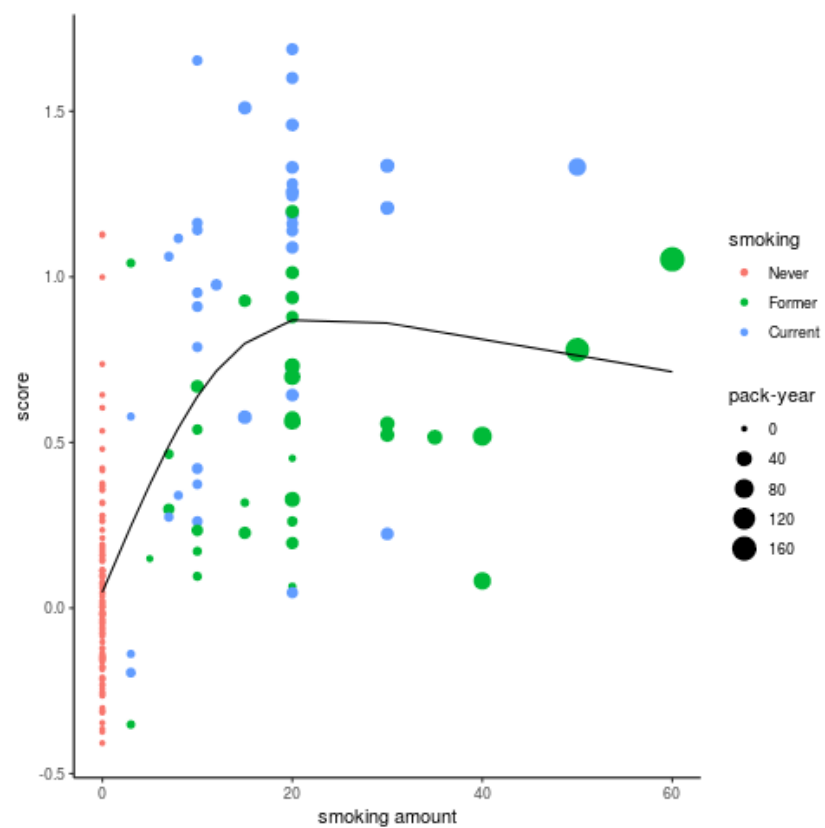
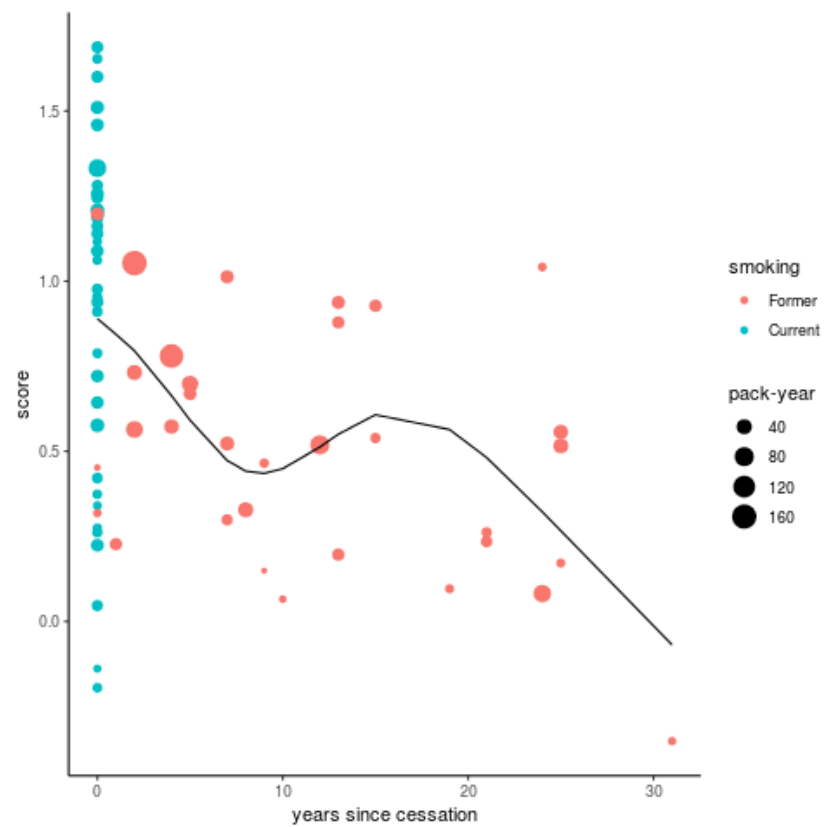


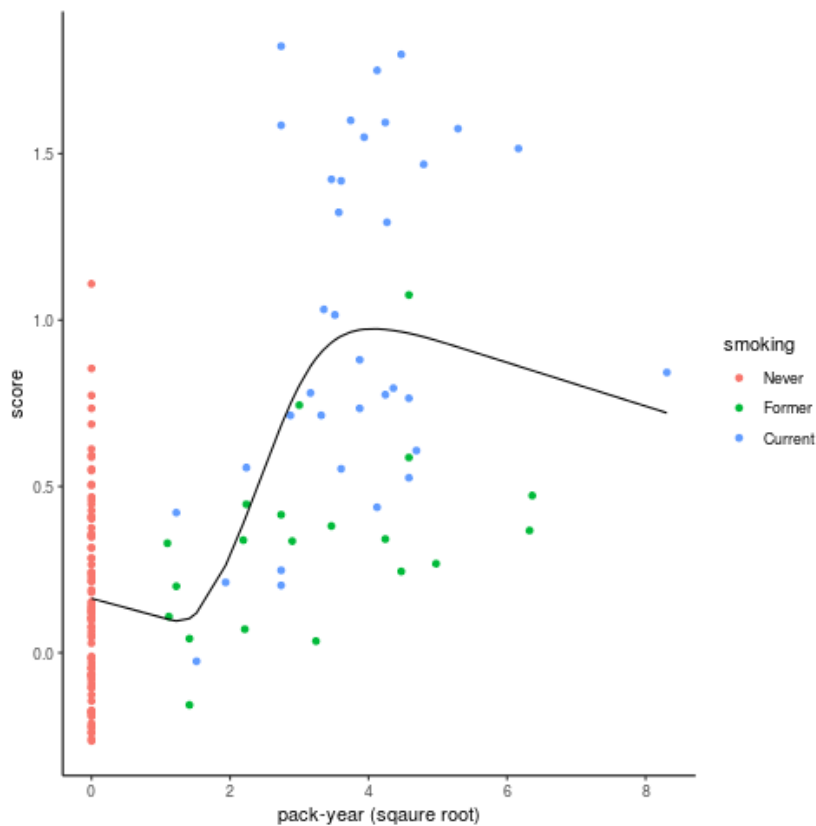
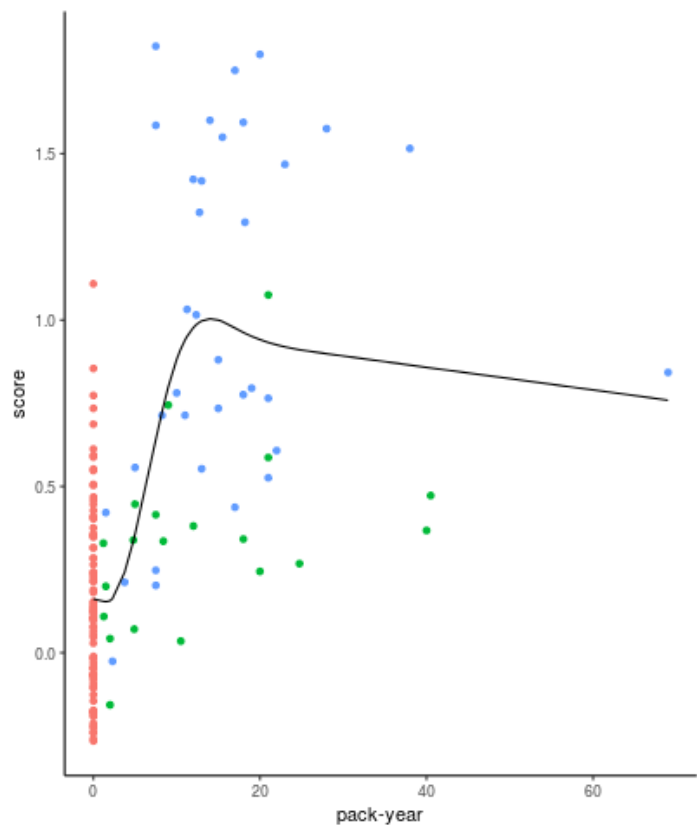
Figure 14. Dose-response relationships between dose of smoking exposure and DNAm score
A. KHT validation set I

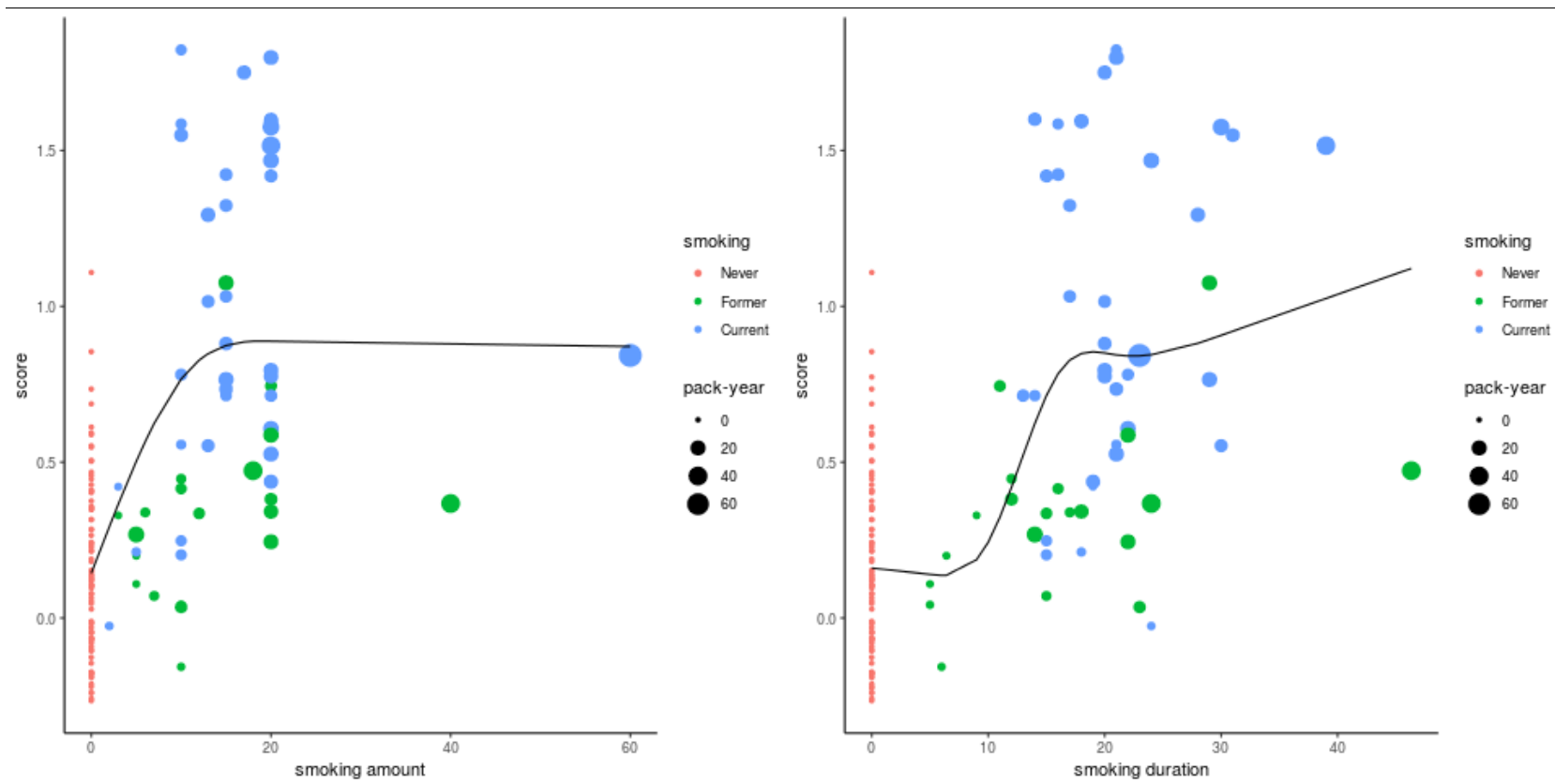


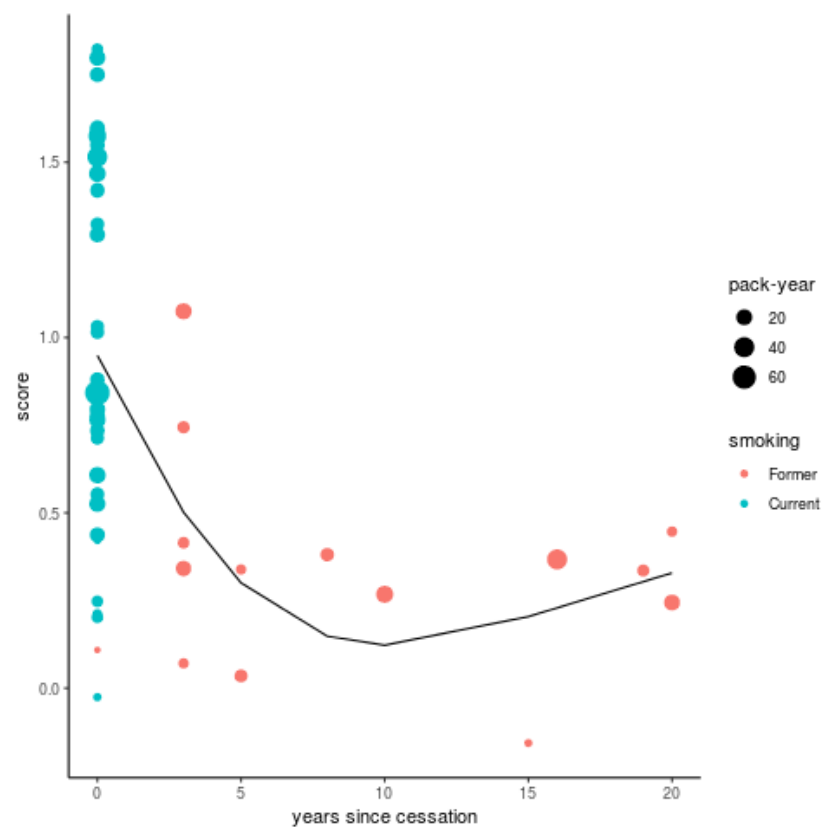




B. KHT validation set II







C. AMDTSS validation set

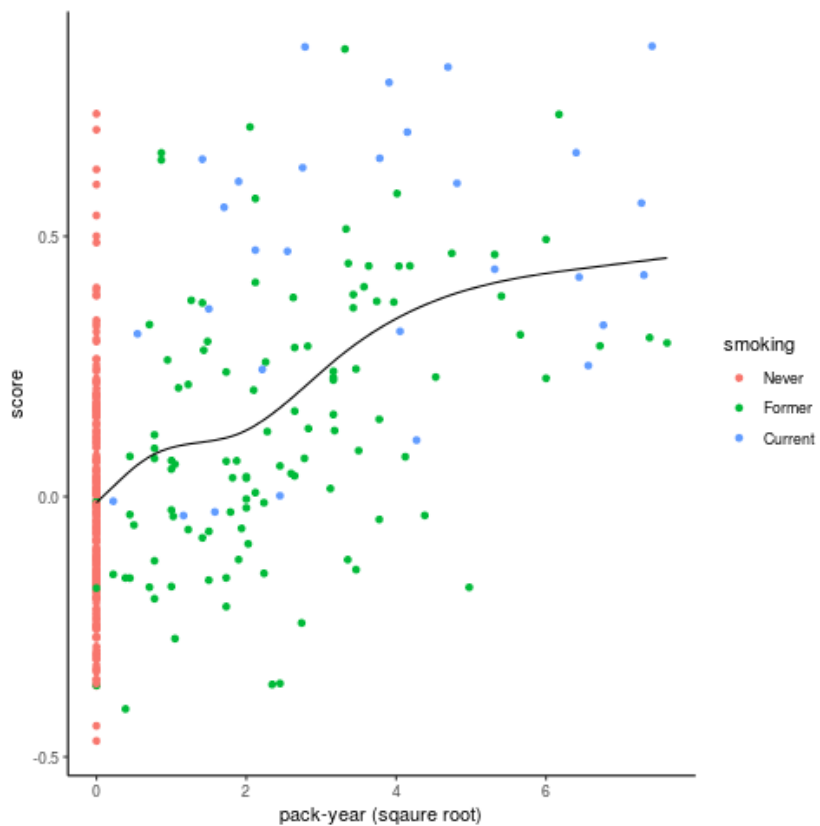
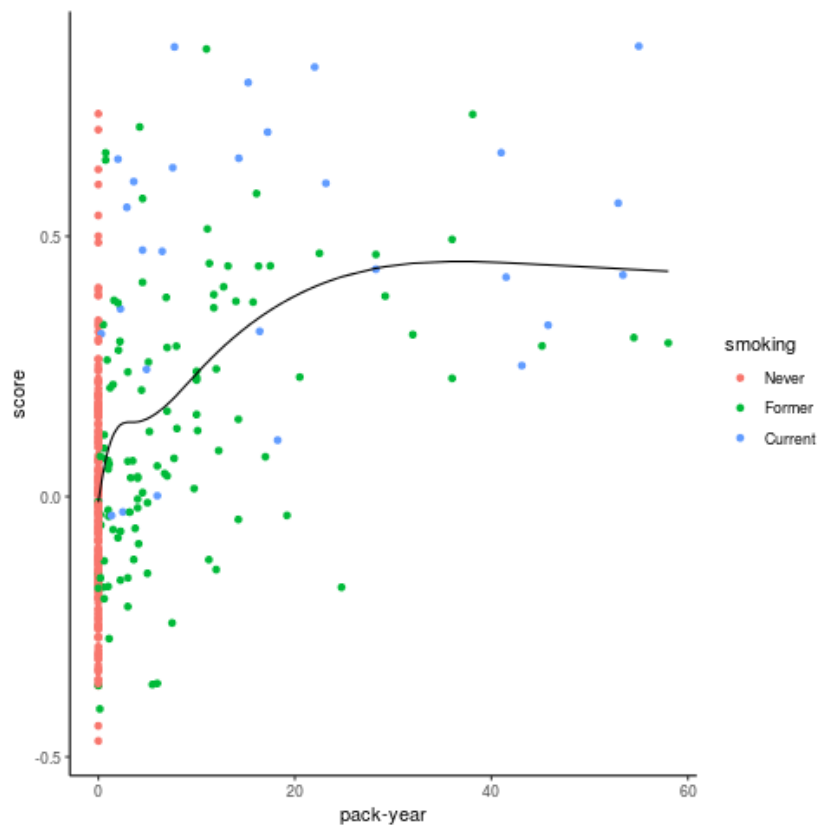
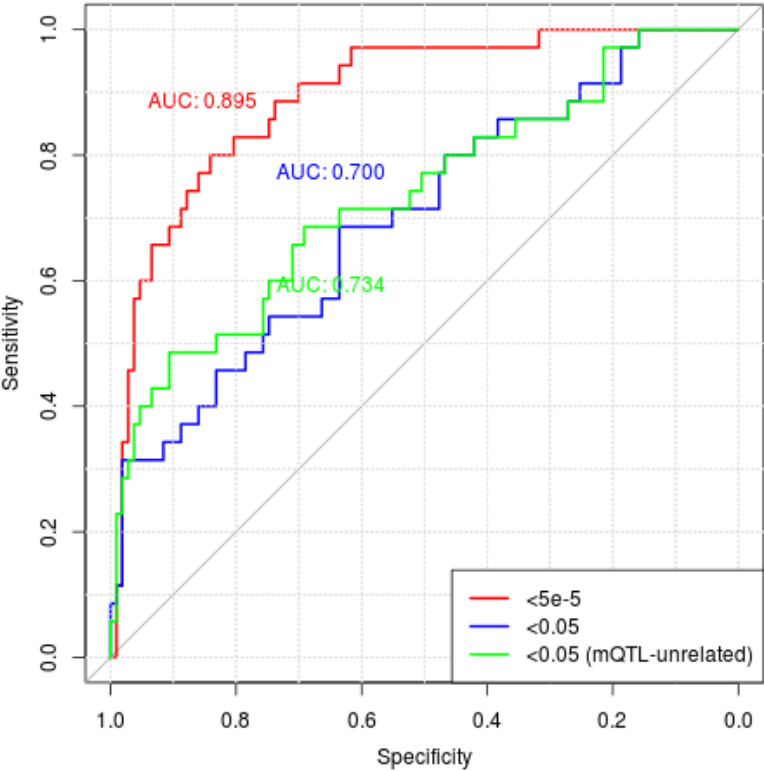
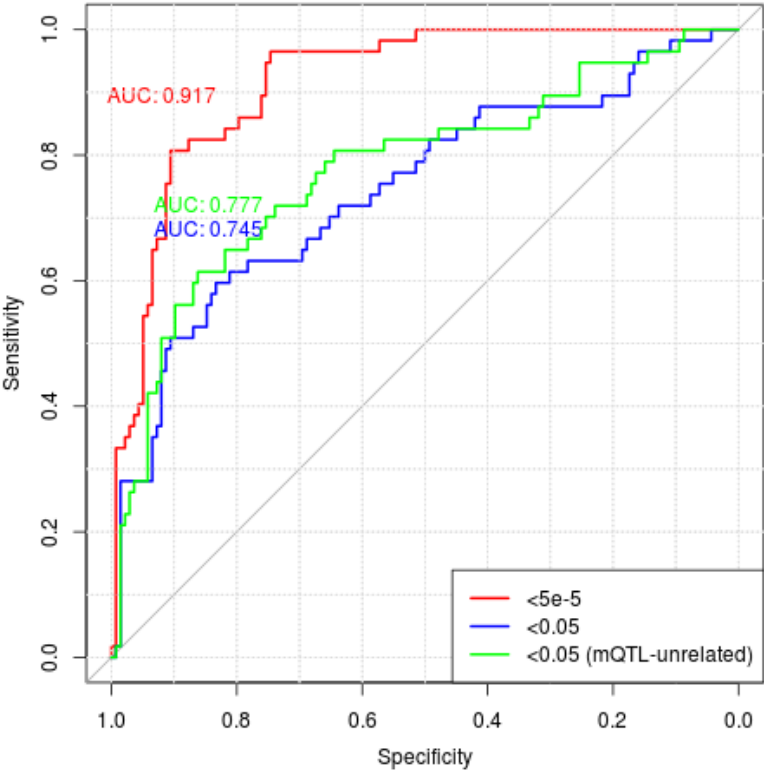


Figure 15. ROC curves for DNAm score as a classifier of ever smoking (Pack-year cutoff=10)
A. KHT validation set I B. KHT validation set II



C. AMDTSS validation set

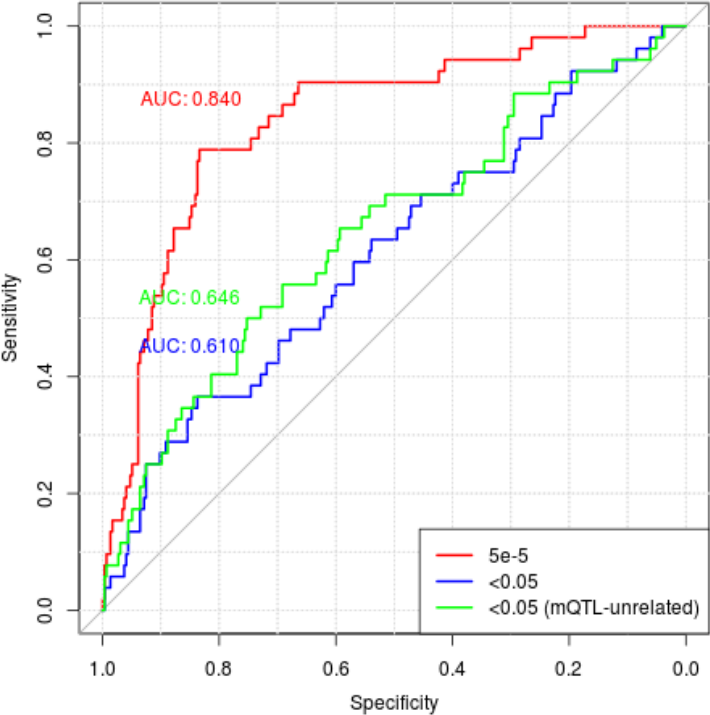
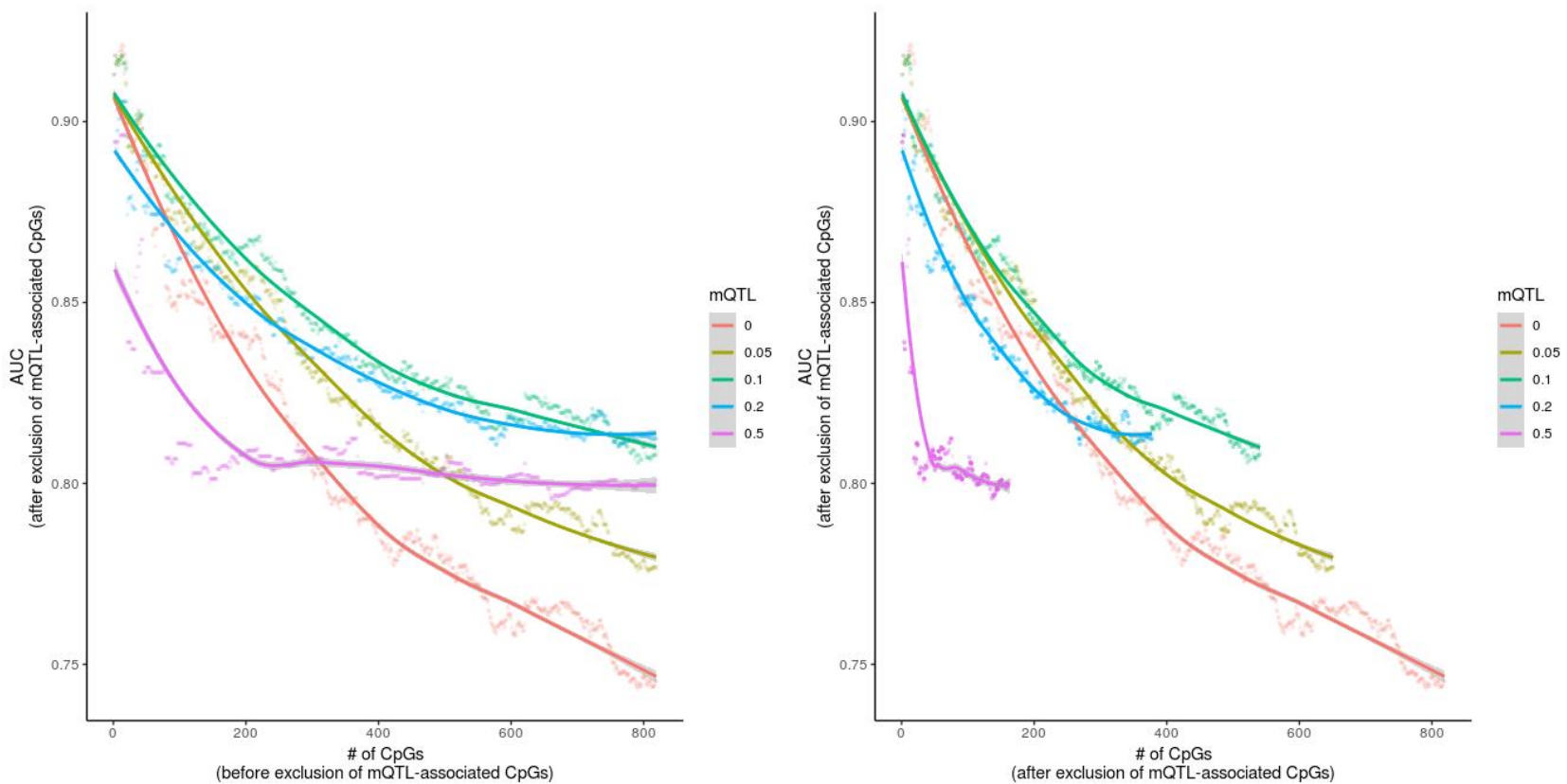
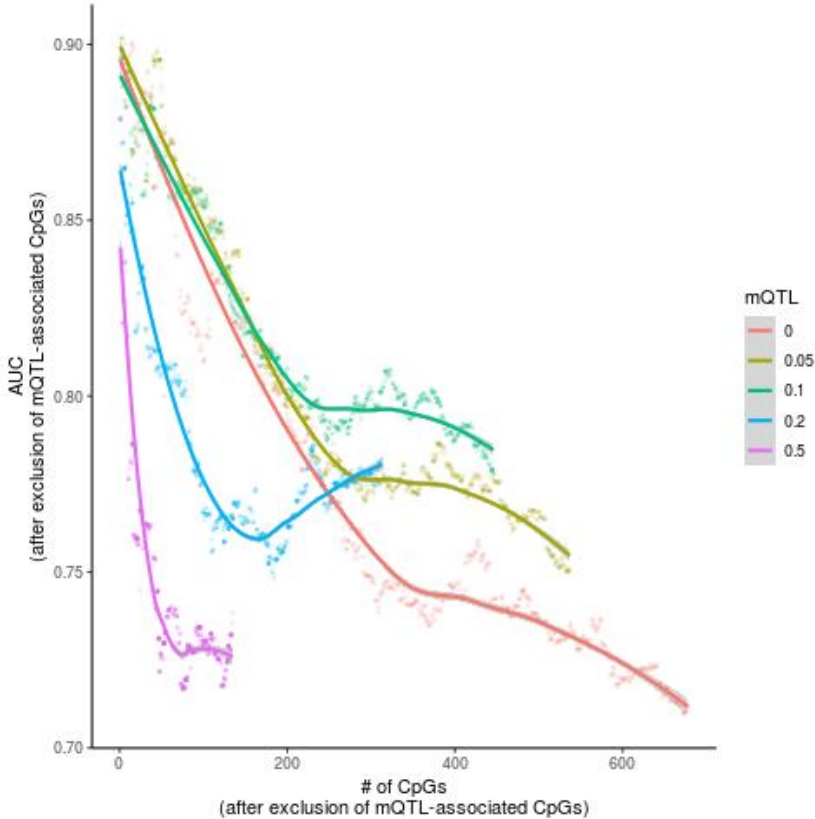
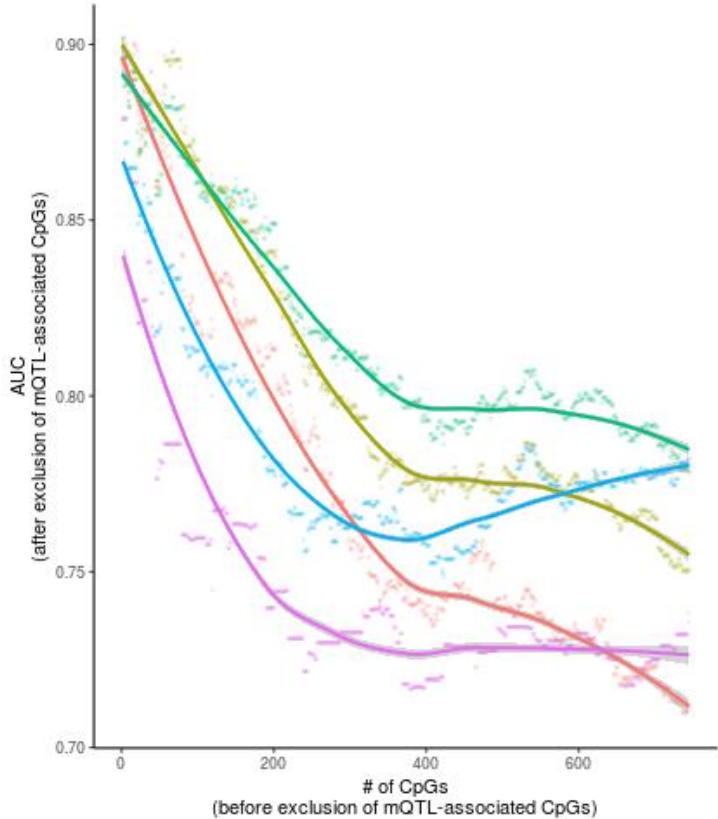


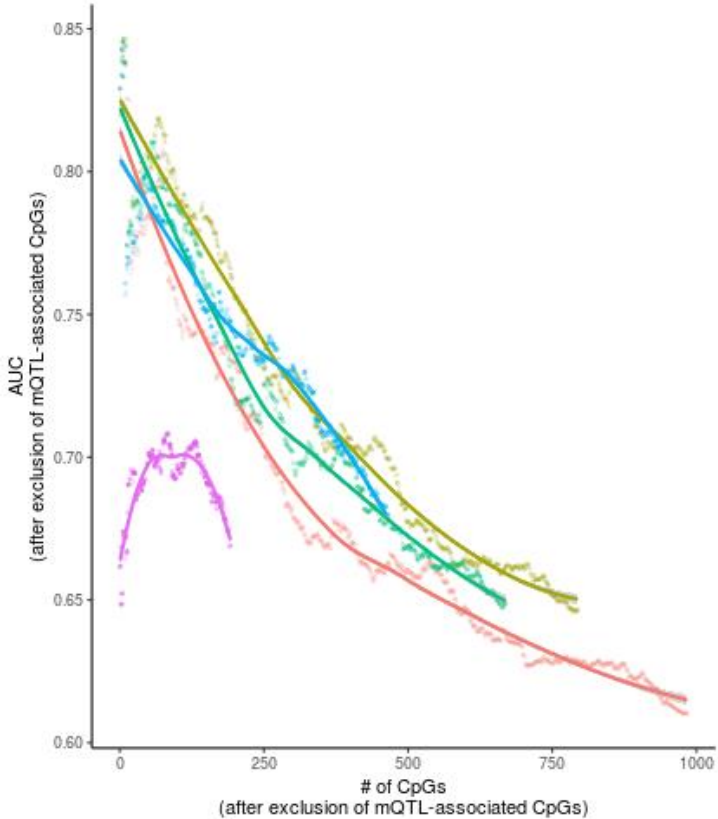
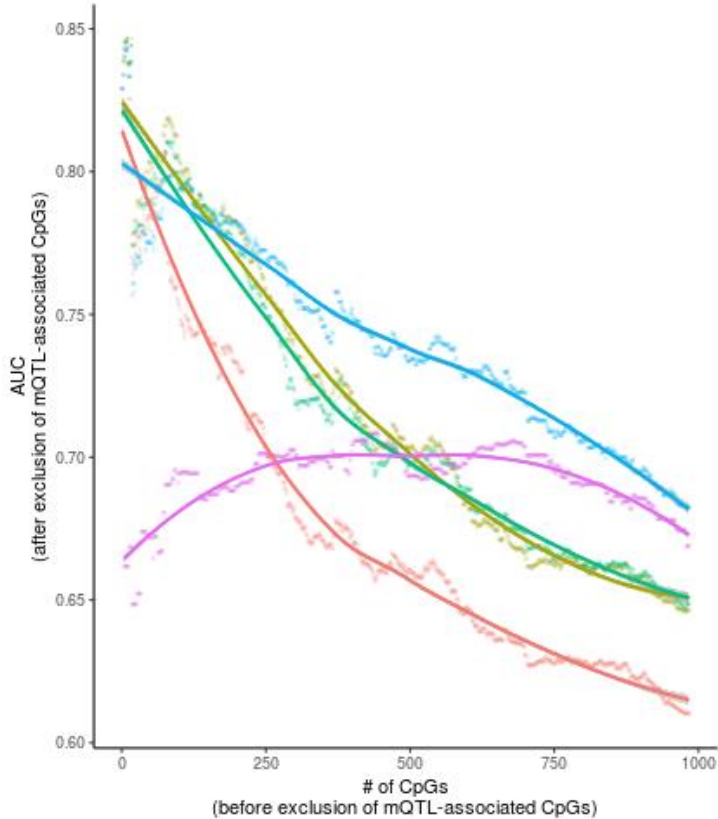
Figure 16. AUC for prediction of smoking using DNAm score of CpGs selected according to the number of top smoking-associated CpGs and exclusion cutoff of mQTLs (KHT validation set I)
A. KHT validation set I



B. KHT validation set II



C. AMDTSS validation set



V. References

1. Waddington CH. The strategy of the genes. Routledge, 2014.
2. Dhingra R, Nwanaji-Enwerem JC, Samet M, Ward-Caviness CK. DNA Methylation Age-Environmental Influences, Health Impacts, and Its Role in Environmental Epidemiology. *Curr Environ Health Rep* 2018.
3. Ho SM, Johnson A, Tarapore P, Janakiram V, Zhang X, Leung YK. Environmental epigenetics and its implication on disease risk and health outcomes. *ILAR J* 2012; 53:289-305.
4. Ehrlich M, Gama-Sosa MA, Huang LH, Midgett RM, Kuo KC, McCune RA, Gehrke C. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic Acids Res* 1982; 10:2709-21.
5. Lister R, Ecker JR. Finding the fifth base: genome-wide sequencing of cytosine methylation. *Genome research* 2009; 19:959-66.
6. Petronis A. Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature* 2010; 465:721.
7. Tobin EW, Goeman JJ, Monajemi R, Gu H, Putter H, Zhang Y, Sliker RC, Stok AP, Thijssen PE, Müller F. DNA methylation signatures link prenatal famine exposure to growth and metabolism. *Nature communications* 2014; 5:5592.
8. Kim E, Kwak SH, Chung HR, Ohn JH, Bae JH, Choi SH, Park KS, Hong J-S, Sung J, Jang HC. DNA methylation profiles in sibling pairs discordant for intrauterine exposure to maternal gestational diabetes. *Epigenetics* 2017; 12:825-32.
9. Joubert BR, Häberg SE, Nilsen RM, Wang X, Vollset SE, Murphy SK, Huang Z, Hoyo C, Midttun Ø, Cupul-Uicab LA. 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. *Environmental health perspectives* 2012; 120:1425.
10. Toledo-Rodriguez M, Lotfipour S, Leonard G, Perron M, Richer L, Veillette S, Pausova Z, Paus T. Maternal smoking during pregnancy is associated with epigenetic modifications of the brain-derived neurotrophic factor-6 exon in adolescent offspring. *Am J Med Genet B Neuropsychiatr Genet* 2010; 153B:1350-4.
11. Salnikow K, Zhitkovich A. Genetic and epigenetic mechanisms in metal carcinogenesis and cocarcinogenesis: nickel, arsenic, and chromium. *Chem Res Toxicol* 2008; 21:28-44.
12. Argos M. Arsenic exposure and epigenetic alterations: recent findings based on the Illumina 450K DNA methylation array. *Current environmental health reports* 2015; 2:137-44.
13. Argos M, Chen L, Jasmine F, Tong L, Pierce BL, Roy S, Paul-Brutus R, Gamble MV, Harper KN, Parvez F. Gene-specific differential DNA methylation and chronic arsenic exposure in an epigenome-wide association study of adults in Bangladesh. *Environmental health perspectives* 2015; 123:64.
14. Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *Am J Hum Genet* 2011; 88:450-7.
15. Zeilinger S, Kuhnel B, Klopp N, Baurecht H, Kleinschmidt A, Gieger C, Weidinger S, Lattka E, Adamski J, Peters A, et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. *PLoS One* 2013; 8:e63812.

16. Gao X, Jia M, Zhang Y, Breitling LP, Brenner H. DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. *Clin Epigenetics* 2015; 7:113.
17. Ambatipudi S, Cuenin C, Hernandez-Vargas H, Ghantous A, Le Calvez-Kelm F, Kaaks R, Barrdahl M, Boeing H, Aleksandrova K, Trichopoulou A, et al. Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. *Epigenomics* 2016; 8:599-618.
18. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, Guan W, Xu T, Elks CE, Aslibekyan S, et al. Epigenetic Signatures of Cigarette Smoking. *Circ Cardiovasc Genet* 2016; 9:436-47.
19. Philibert R, Plume JM, Gibbons FX, Brody GH, Beach S. The impact of recent alcohol use on genome wide DNA methylation signatures. *Frontiers in genetics* 2012; 3:54.
20. Liu C, Marioni R, Hedman ÅK, Pfeiffer L, Tsai P, Reynolds L, Just A, Duan Q, Boer C, Tanaka T. A DNA methylation biomarker of alcohol consumption. *Molecular psychiatry* 2016.
21. Berkel TD, Pandey SC. Emerging role of epigenetic mechanisms in alcohol addiction. *Alcoholism: Clinical and Experimental Research* 2017; 41:666-80.
22. Wahl S, Drong A, Lehne B, Loh M, Scott WR, Kunze S, Tsai P-C, Ried JS, Zhang W, Yang Y. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* 2017; 541:81.
23. Fasanelli F, Baglietto L, Ponzi E, Guida F, Campanella G, Johansson M, Grankvist K, Johansson M, Assumma MB, Naccarati A, et al. Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. *Nat Commun* 2015; 6:10192.
24. Godfrey KM, Costello PM, Lillycrop KA. The developmental environment, epigenetic biomarkers and long-term health. *J Dev Orig Health Dis* 2015; 6:399-406.
25. Lemire M, Zaidi SH, Ban M, Ge B, Aissi D, Germain M, Kassam I, Wang M, Zanke BW, Gagnon F, et al. Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci. *Nat Commun* 2015; 6:6326.
26. Volkov P, Olsson AH, Gillberg L, Jorgensen SW, Brons C, Eriksson KF, Groop L, Jansson PA, Nilsson E, Ronn T, et al. A Genome-Wide mQTL Analysis in Human Adipose Tissue Identifies Genetic Variants Associated with DNA Methylation, Gene Expression and Metabolic Traits. *PLoS One* 2016; 11:e0157776.
27. Gao X, Thomsen H, Zhang Y, Breitling LP, Brenner H. The impact of methylation quantitative trait loci (mQTLs) on active smoking-related DNA methylation changes. *Clin Epigenetics* 2017; 9:87.
28. Dupont J-M, Tost J, Jammes H, Gut IG. De novo quantitative bisulfite sequencing using the pyrosequencing technology. *Analytical biochemistry* 2004; 333:119-27.
29. Eads CA, Danenberg KD, Kawakami K, Saltz LB, Blake C, Shibata D, Danenberg PV, Laird PW. MethyLight: a high-throughput assay to measure DNA methylation. *Nucleic acids research* 2000; 28:e32-00.
30. Ehrich M, Nelson MR, Stanssens P, Zabeau M, Liloglou T, Xinarianos G, Cantor CR, Field JK, van den Boom D. Quantitative high-throughput analysis of DNA methylation patterns by base-specific cleavage and mass spectrometry. *Proceedings of the National Academy of Sciences* 2005; 102:15785-90.
31. Bell JT, Saffery R. The value of twins in epigenetic epidemiology. *Int J Epidemiol* 2012; 41:140-50.
32. Castillo-Fernandez JE, Spector TD, Bell JT. Epigenetics of discordant monozygotic twins: implications for disease. *Genome Med* 2014; 6:60.

33. Huang YT, Chu S, Loucks EB, Lin CL, Eaton CB, Buka SL, Kelsey KT. Epigenome-wide profiling of DNA methylation in paired samples of adipose tissue and blood. *Epigenetics* 2016; 11:227-36.
34. Walton E, Hass J, Liu J, Roffman JL, Bernardoni F, Roessner V, Kirsch M, Schackert G, Calhoun V, Ehrlich S. Correspondence of DNA Methylation Between Blood and Brain Tissue and Its Application to Schizophrenia Research. *Schizophr Bull* 2016; 42:406-14.
35. Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. *Biostatistics* 2012; 13:539-52.
36. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics* 2012; 13:86.
37. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics* 2007; 3:e161.
38. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods* 2015; 12:453.
39. Teschendorff AE, Zhuang J, Widschwendter M. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics* 2011; 27:1496-505.
40. Zou J, Lippert C, Heckerman D, Aryee M, Listgarten J. Epigenome-wide association studies without the need for cell-type composition. *Nature methods* 2014; 11:309.
41. Collaborators GBDT. Smoking prevalence and attributable disease burden in 195 countries and territories, 1990-2015: a systematic analysis from the Global Burden of Disease Study 2015. *Lancet* 2017; 389:1885-906.
42. Park S, Jee SH, Shin HR, Park EH, Shin A, Jung KW, Hwang SS, Cha ES, Yun YH, Park SK, et al. Attributable fraction of tobacco smoking on cancer using population-based nationwide cancer incidence and mortality data in Korea. *BMC Cancer* 2014; 14:406.
43. Caraballo RS, Giovino GA, Pechacek TF, Mowery PD. Factors associated with discrepancies between self-reports on cigarette smoking and measured serum cotinine levels among persons aged 17 years or older: Third National Health and Nutrition Examination Survey, 1988-1994. *Am J Epidemiol* 2001; 153:807-14.
44. Vartiainen E, Seppala T, Lillsunde P, Puska P. Validation of self reported smoking by serum cotinine measurement in a community-based study. *J Epidemiol Community Health* 2002; 56:167-70.
45. Kvalvik LG, Nilsen RM, Skjaerven R, Vollset SE, Midttun O, Ueland PM, Haug K. Self-reported smoking status and plasma cotinine concentrations among pregnant women in the Norwegian Mother and Child Cohort Study. *Pediatr Res* 2012; 72:101-7.
46. Park MB, Kim CB, Nam EW, Hong KS. Does South Korea have hidden female smokers: discrepancies in smoking rates between self-reports and urinary cotinine level. *BMC Womens Health* 2014; 14:156.
47. Florescu A, Ferrence R, Einarson T, Selby P, Soldin O, Koren G. Methods for quantification of exposure to cigarette smoking and environmental tobacco smoke: focus on developmental toxicology. *Ther Drug Monit* 2009; 31:14-30.
48. Britton GR, Brinthaup J, Stehle JM, James GD. Comparison of self-reported smoking and urinary cotinine levels in a rural pregnant population. *J Obstet Gynecol Neonatal Nurs* 2004; 33:306-11.
49. Caraballo RS, Giovino GA, Pechacek TF. Self-reported cigarette smoking vs. serum cotinine among U.S. adolescents. *Nicotine Tob Res* 2004; 6:19-25.

50. Kandel DB, Schaffran C, Griesler PC, Hu M-C, Davies M, Benowitz N. Salivary cotinine concentration versus self-reported cigarette smoking: three patterns of inconsistency in adolescence. *Nicotine & tobacco research* 2006; 8:525-37.
51. Lee KW, Pausova Z. Cigarette smoking and DNA methylation. *Front Genet* 2013; 4:132.
52. Watanabe T, Imoto I, Kosugi Y, Fukuda Y, Mimura J, Fujii Y, Isaka K, Takayama M, Sato A, Inazawa J. Human arylhydrocarbon receptor repressor (AHRR) gene: genomic structure and analysis of polymorphism in endometriosis. *J Hum Genet* 2001; 46:342-6.
53. Andersen AM, Philibert RA, Gibbons FX, Simons RL, Long J. Accuracy and utility of an epigenetic biomarker for smoking in populations with varying rates of false self-report. *Am J Med Genet B Neuropsychiatr Genet* 2017; 174:641-50.
54. Shenker NS, Ueland PM, Polidoro S, van Veldhoven K, Ricceri F, Brown R, Flanagan JM, Vineis P. DNA methylation as a long-term biomarker of exposure to tobacco smoke. *Epidemiology* 2013; 24:712-6.
55. Zhang Y, Florath I, Saum KU, Brenner H. Self-reported smoking, serum cotinine, and blood DNA methylation. *Environ Res* 2016; 146:395-403.
56. Philibert R, Hollenbeck N, Andersen E, Osborn T, Gerrard M, Gibbons FX, Wang K. A quantitative epigenetic approach for the assessment of cigarette consumption. *Front Psychol* 2015; 6:656.
57. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. *Bioinformatics* 2008; 24:1547-8.
58. Triche TJ, Jr., Weisenberger DJ, Van Den Berg D, Laird PW, Siegmund KD. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res* 2013; 41:e90.
59. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 2014; 30:1363-9.
60. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. Evaluation of the Infinium Methylation 450K technology. *Epigenomics* 2011; 3:771-84.
61. Maksimovic J, Gordon L, Oshlack A. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome Biol* 2012; 13:R44.
62. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, Beck S. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* 2013; 29:189-96.
63. Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, Reinius L, Acevedo N, Taub M, Ronninger M. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature biotechnology* 2013; 31:142.
64. Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome biology* 2014; 15:R31.
65. Houseman EA, Molitor J, Marsit CJ. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics* 2014; 30:1431-9.
66. Teschendorff AE, Relton CL. Statistical and integrative system-level analysis of DNA methylation data. *Nat Rev Genet* 2018; 19:129-47.
67. Sung J, Cho SI, Lee K, Ha M, Choi EY, Choi JS, Kim H, Kim J, Hong KS, Kim Y, et al. Healthy Twin: a twin-family study of Korea--protocols and current status. *Twin Res Hum Genet* 2006; 9:844-8.

68. Gombojav B, Song YM, Lee K, Yang S, Kho M, Hwang YC, Ko G, Sung J. The Healthy Twin Study, Korea updates: resources for omics and genome epidemiology studies. *Twin Res Hum Genet* 2013; 16:241-5.
69. Odefrey F, Stone J, Gurrin LC, Byrnes GB, Apicella C, Dite GS, Cawson JN, Giles GG, Treloar SA, English DR, et al. Common genetic variants associated with breast cancer and mammographic density measures that predict disease. *Cancer Res* 2010; 70:1449-58.
70. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007; 8:118-27.
71. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 2012; 28:882-3.
72. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; 81:559-75.
73. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology* 2004; 3:1-25.
74. Peters TJ, Buckley MJ, Statham AL, Pidsley R, Samaras K, Lord RV, Clark SJ, Molloy PL. De novo identification of differentially methylated regions in the human genome. *Epigenetics & chromatin* 2015; 8:6.
75. Cooper H, Hedges LV, Valentine JC. The handbook of research synthesis and meta-analysis. Russell Sage Foundation, 2009.
76. Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics* 2005; 30:261-93.
77. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software* 2010; 36:1-48.
78. Desquilbet L, Mariotti F. Dose-response analyses using restricted cubic spline functions in public health research. *Stat Med* 2010; 29:1037-57.
79. Harrell Jr FE, Harrell Jr MFE, Hmisc D. Package 'rms'. Vanderbilt University 2019:229.
80. Harrell Jr FE, Harrell Jr MFE. Package 'Hmisc'. CRAN2018 2019:235-6.
81. Li M, Zou D, Li Z, Gao R, Sang J, Zhang Y, Li R, Xia L, Zhang T, Niu G, et al. EWAS Atlas: a curated knowledgebase of epigenome-wide association studies. *Nucleic Acids Res* 2019; 47:D983-D8.
82. Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, Zhukova A, Brazma A, Parkinson H. Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* 2010; 26:1112-8.
83. Aulchenko YS, Ripke S, Isaacs A, Van Duijn CM. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* 2007; 23:1294-6.
84. Shenker NS, Polidoro S, van Veldhoven K, Sacerdote C, Ricceri F, Birrell MA, Belvisi MG, Brown R, Vineis P, Flanagan JM. Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. *Human molecular genetics* 2012; 22:843-51.
85. Dua P, Kang HS, Hong SM, Tsao MS, Kim S, Lee DK. Alkaline phosphatase ALPPL-2 is a novel pancreatic carcinoma-associated protein. *Cancer Res* 2013; 73:1934-45.
86. Zhang Y, Yang R, Burwinkel B, Breitling LP, Holleczeck B, Schottker B, Brenner H. F2RL3 methylation in blood DNA is a strong predictor of mortality. *Int J Epidemiol* 2014; 43:1215-25.

87. Gerard A, Patino-Lopez G, Beemiller P, Nambiar R, Ben-Aissa K, Liu Y, Totah FJ, Tyska MJ, Shaw S, Krummel MF. Detection of rare antigen-presenting cells through T cell-intrinsic meandering motility, mediated by Myo1g. *Cell* 2014; 158:492-505.
88. Diniz MG, Franca JA, Vilas-Boas FAS, de Souza FTA, Calin GA, Gomez RS, de Sousa SF, Gomes CC. The long noncoding RNA KIAA0125 is upregulated in ameloblastomas. *Pathol Res Pract* 2019; 215:466-9.
89. Philibert R, Hollenbeck N, Andersen E, McElroy S, Wilson S, Vercande K, Beach SR, Osborn T, Gerrard M, Gibbons FX, et al. Reversion of AHRR Demethylation Is a Quantitative Biomarker of Smoking Cessation. *Front Psychiatry* 2016; 7:55.
90. Tsaprouni LG, Yang TP, Bell J, Dick KJ, Kanoni S, Nisbet J, Vinuela A, Grundberg E, Nelson CP, Meduri E, et al. Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. *Epigenetics* 2014; 9:1382-96.
91. The Health Consequences of Smoking-50 Years of Progress: A Report of the Surgeon General. Atlanta (GA), 2014.
92. Mons U, Muezzinler A, Gellert C, Schottker B, Abnet CC, Bobak M, de Groot L, Freedman ND, Jansen E, Kee F, et al. Impact of smoking and smoking cessation on cardiovascular events and mortality among older adults: meta-analysis of individual participant data from prospective cohort studies of the CHANCES consortium. *BMJ* 2015; 350:h1551.
93. Giuliani C, Biggs D, Nguyen TT, Marasco E, De Fanti S, Garagnani P, Le Phan MT, Nguyen VN, Luiselli D, Romeo G. First evidence of association between past environmental exposure to dioxin and DNA methylation of CYP1A1 and IGF2 genes in present day Vietnamese population. *Environ Pollut* 2018; 242:976-85.
94. Alhamdow A, Lindh C, Hagberg J, Graff P, Westberg H, Kraus AM, Albin M, Gustavsson P, Tinnerberg H, Broberg K. DNA methylation of the cancer-related genes F2RL3 and AHRR is associated with occupational exposure to polycyclic aromatic hydrocarbons. *Carcinogenesis* 2018; 39:869-78.
95. Kipler M, Engstrom K, Mlakar SJ, Bottai M, Ahmed S, Hossain MB, Raqib R, Vahter M, Broberg K. Sex-specific effects of early life cadmium exposure on DNA methylation and implications for birth weight. *Epigenetics* 2013; 8:494-503.
96. Ladd-Acosta C. Epigenetic Signatures as Biomarkers of Exposure. *Curr Environ Health Rep* 2015; 2:117-25.
97. Tsai P-C, Bell JT. Power and sample size estimation for epigenome-wide association scans to detect differential DNA methylation. *International journal of epidemiology* 2015; 44:1429-41.
98. Elliott HR, Tillin T, McArdle WL, Ho K, Duggirala A, Frayling TM, Davey Smith G, Hughes AD, Chaturvedi N, Relton CL. Differences in smoking associated DNA methylation patterns in South Asians and Europeans. *Clin Epigenetics* 2014; 6:4.
99. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; 143:29-36.
100. Youden WJ. Index for rating diagnostic tests. *Cancer* 1950; 3:32-5.
101. Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 2005; 47:458-72.
102. Hand DJ, Till RJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine learning* 2001; 45:171-86.

103. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics* 2011; 12:77.
104. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; 44:837-45.
105. Sun X, Xu W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters* 2014; 21:1389-93.
106. Pencina MJ, D'Agostino Sr RB, D'Agostino Jr RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in medicine* 2008; 27:157-72.
107. Kundu S, Aulchenko YS, van Duijn CM, Janssens ACJ. PredictABEL: an R package for the assessment of risk prediction models. *European journal of epidemiology* 2011; 26:261.
108. Shmueli G. To explain or to predict? *Statistical science* 2010; 25:289-310.
109. Halligan S, Altman DG, Mallett S. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. *Eur Radiol* 2015; 25:932-9.

VI. Abstract in Korean (국문 초록)

흡연 관련 후성유전학 지표 발굴 연구

서울대학교 보건대학원
보건학과 유전체역학 전공
김 은 애

흡연은 폐/심혈관계 질환 및 폐암 등 여러 질환들에 대한 교정 가능한 (modifiable) 위험 요인임에도 불구하고, 흡연의 기여 사망률은 전세계적으로 11.5 % 에 달한다. 이러한 흡연 관련 건강영향을 평가하기 위해서는 정확한 흡연 노출 및 노출량 측정이 선행되어야 한다. 흡연 노출 평가에 이용되는 대표적인 방법에는 자가 보고 (self-report) 기반 설문 도구와 더불어, 코티닌 (cotinine) 과 같은 체내의 흡연 관련 대사체의 농도를 측정하는 등 생체 지표 (biomarker) 들을 이용하는 방법이 있으나, 최근 흡연 노출만을 제한적으로 반영한다는 한계점을 가진다. 이에 따라, 현재 뿐만 아니라 과거 흡연 노출을

반영하는 지속성 및 안정성을 보이는 지표들을 발굴하기 위하여 흡연 관련 후성유전 연구가 활발히 진행되고 있다.

후성유전 (Epigenetics) 은 DNA 염기서열 상의 변화 없이 유전자의 발현에 영향을 주는 유전적 현상을 가리키며, DNA 메틸화 (DNA methylation) 는 유전적인 요인 뿐만 아니라 생애 전반에 걸쳐 노출되는 여러가지 환경적인 요인들에 의해서 결정되는 대표적인 후성유전학적 지표이다. DNA 메틸화는 가변적인 특성 때문에 특정한 환경적 요인에 의한 특이적인 DNA 메틸화 변화를 발굴하기 위해서는 여러 가지 잠재 교란 요인들이 통제되어야 한다. 이에 따라, 본 연구는 유전 및 환경적 요인에 의한 교호 작용을 통제할 수 있는 일란성 쌍둥이 및 그 직계 가족들의 샘플을 이용하여 흡연 노출에 대해 특이적으로 변화하는 DNA 메틸화 지표를 발굴하고자 수행되었다. 먼저, 전장 후성유전체 연관 분석을 통해 흡연 여부에 따른 쌍둥이 간의 DNA 메틸화 차이를 관찰하는 분석을 수행하였다. 나아가, DNA 메틸화 수준의 변화와 연관된 단일염기성다형성 (Single nucleotide polymorphisms, SNP) 변이를 찾는 mQTL (methylation quantitative loci) 분석을 수행하였다. 최종적으로, 연관 분석에서 발굴된 흡연 관련 DNA 메틸화 지표들을 바탕으로 검증 데이터 (validation set) 의 각 샘플들에 대하여 DNA 메틸화 점수를 부여하여, DNA 메틸화 기반 점수의 흡연 예측 능력을 평가하였다.

DNA 메틸화 분석을 위해 한국인 가족-쌍둥이 (KHT) 코호트 및 호주의 Australian Mammographic Density Twins and Sisters Study (AMDTSS) 코호트로 부터 각각 534 명 및 132 명의 대상자들이 포함되었다. KHT 코호트와 AMDTSS 코호트에서 각각 156 쌍, 66 쌍의 일란성 쌍둥이

대상자들이 분석에 포함되었다. 말초 혈액 백혈구 샘플에서 DNA 를 추출한 다음, KHT 의 385 명의 대상자 및 AMDTSS 의 모든 대상자들의 샘플은 Illumina 사의 Infinium HumanMethylation 450 BeadChip 로 어레이하여 유전체 내 약 450,000 개 이상의 DNA 메틸화 위치에 대한 DNA 메틸화 수준의 데이터를 얻었으며, 총 149 명의 KHT 코호트의 샘플들은 Illumina 사의 Infinium MethylationEPIC BeadChip 로 어레이하여 유전체 내 총 850,000 개 이상의 위치에 대한 DNA 메틸화 정도를 측정하였다. R 소프트웨어의 생물정보학 관련 패키지들을 이용하여 기존 대규모 메타 연구에서 밝혀진 18,000 개 가량의 흡연 관련 DNA 메틸화 지표에 대해 일관성 쌍둥이 내의 DNA 메틸화의 차이를 평가하고, 나아가 KHT 및 AMDTSS 코호트의 결과로 메타 분석을 수행하였다. 또한, 약 18,000 개의 각 DNA 메틸화 지표에 대해서 $\pm 1\text{Mb}$ 위치 내의 SNP 과의 연관성을 검정하는 mQTL 분석을 수행하였다. 나아가, DNA 메틸화 점수는 크게 다음과 같이 총 3 가지 모형에 대한 흡연 예측 능력을 평가 및 비교하였다. (1) 유의 수준 5×10^{-5} 미만의 흡연과의 연관성을 보인 DNA 메틸화 지표들로 구성된 세트, (2) 유의 수준 0.05 미만의 DNA 메틸화 지표 중 mQTL 의 영향을 받는 지표들을 제거한 나머지 지표들로 구성된 세트 및 (3) mQTL 의 영향을 고려하지 않은 세트에 대해 DNA 메틸화 점수를 계산하여 흡연 여부에 대한 예측력을 평가하였다.

후성유전체 연관 메타 분석에 따르면, *AHRR* (cg23576855), *ALPPL2* (cg21566642, cg01940273, cg05951221), *MYOIG* (cg12803068) 와 *F2RL3* (cg03636183) 등의 유전자좌 내의 CpG 위치에서 DNA 메틸화 수준의 차이가

관찰되었다. mQTL 분석에서는 기존 연구에서 밝혀진 흡연 관련 DNA 메틸화 지표 중 약 19.6%가 적어도 하나의 근위 단일염기성다형성과 연관이 있는 것으로 확인되었다. 상위 연관 지표들로 계산된 DNA 메틸화 점수는 흡연 여부에 대한 예측력 (AUC) 은 검증 데이터 세트에 따라 약 0.84~0.92 으로 계산되었다. 유의 수준 0.05 미만의 지표 중 mQTL 의 영향을 받는 DNA 메틸화 지표들을 제거한 세트의 AUC 는 0.65~0.78, mQTL 과의 연관성을 고려하지 않은 세트의 AUC 는 0.61~0.75 에 비해 통계적으로 유의한 수준으로 높았다.

본 연구는 일란성 쌍둥이 및 가족 연구를 바탕으로 기존에 알려진 흡연 관련 후성유전학 지표들 중에 유전적 변이를 받는 지표들과 흡연 노출을 특이적으로 반영하는 지표들을 구분하고, 각 지표들의 흡연에 대한 예측 성능을 비교하였다. 나아가, 흡연 관련 DNA 메틸화 지표는 정확하게 흡연 노출력을 평가하고, 흡연 관련 건강 영향 평가에 활용될 것으로 기대된다.

주요어: 후성유전학, DNA 메틸화, 후성유전체연관분석, 흡연, 생체 지표, 쌍둥이 연구

학번: 2015-31282